

Isabel Alvarez · Mario Arias-Oliva ·
Adrian-Horia Dediu · Nuno Silva (Eds.)

LNCS 15939

Ethical and Social Impacts of Information and Communication Technology

22nd International Conference, ETHICOMP 2025
Lisbon, Portugal, September 17–19, 2025
Proceedings

Lecture Notes in Computer Science

15939

Founding Editors

Gerhard Goos

Juris Hartmanis

Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*

Wen Gao, *Peking University, Beijing, China*

Bernhard Steffen , *TU Dortmund University, Dortmund, Germany*

Moti Yung , *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Isabel Alvarez · Mario Arias-Oliva ·
Adrian-Horia Dediu · Nuno Silva
Editors

Ethical and Social Impacts of Information and Communication Technology

22nd International Conference, ETHICOMP 2025
Lisbon, Portugal, September 17–19, 2025
Proceedings

Editors

Isabel Alvarez 
Lusfada University
Lisbon, Portugal

Adrian-Horia Dediu 
Universidade Autónoma de Lisboa
Lisbon, Portugal

Mario Arias-Oliva 
Complutense University of Madrid
Madrid, Spain

Nuno Silva 
Lusfada University
Lisbon, Portugal

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-032-01428-3

ISBN 978-3-032-01429-0 (eBook)

<https://doi.org/10.1007/978-3-032-01429-0>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2026

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

This volume contains the full papers presented at **ETHICOMP 2025: 22nd International Conference on the Ethical and Social Impacts of ICT**, held on September 17–19, 2025, in Lisbon, Portugal.

ETHICOMP is a well-established international conference series that explores the ethical and social implications of information and communication technologies (ICTs).

In 1994, Professor Simon Rogerson devised the word **ETHICOMP**—which stands for *ETHical COMPuting*—to designate the social and ethical dimensions of information technology. The ETHICOMP conference series was launched at De Montfort University (DMU) in 1995 by the Centre for Computing and Social Responsibility (CCSR). Simon Rogerson and Terry Bynum were the co-founders and joint directors until 2013, when they both stepped down and handed over the conference to the ETHICOMP community. Sustained by the founders’ vision, the enthusiasm of successive generations of participants, and the clear identity that the ETHICOMP name provides, the 2025 edition celebrated the latest research into the spectacular advances of artificial intelligence, as well as the **thirtieth** anniversary of a community dedicated to responsible computing¹.

Traditionally, the conference is held approximately every 18 months; previous editions have been held in cities such as *Logroño, Spain (2024)*, *Turku, Finland (2022)*, *Bologna, Italy (2020)*, and *Paris, France (2018)*, among others.

ETHICOMP 2025 was a *multitrack conference* that united work on ethics, technology, society, and education. Authors initially submitted to ten named tracks and an open track. For these proceedings, we have merged kindred tracks into five coherent chapters, each absorbing one or more original labels. The result preserves every track’s identity while delivering a balanced progression from foundational ethics through societal questions and education to concrete systems and cybersecurity.

Unified Theme	Original Call-for-Papers Track Titles
Ethics: AI Ethics, Governance and Values	<i>Artificial Intelligence: Interpretability and Measurement of its Underlying Ethics • How Is AI Changing Us? How Does It Change the World? • Ethical Challenges for Business in a Digital World • Values: The Basis for Responsible Technology and Digital Interactions</i>

(continued)

¹ A podcast interview recorded on 16 September 2022—“Re-imagining Ethical Digital Technology,” produced by Decentered Media (<https://decentered.co.uk/leicester-smart-city-profesor-simon-rogereson-ethical-digital-technologies/>)— and its companion video, which includes the event slides (<https://youtu.be/ZEpAS2DYrUU>), offer additional insights into Professor Rogerson’s foundational work.

(continued)

Society: Social Networks, Youth and Gender in Digital Society	<i>AI, Social Networks, and Their Influence on Youth • Gender and Future Digital Technologies • (De)stigmatizing Digital Interactions and Technologies</i>
Education: Learning and Literacy in the AI Era	<i>Ethical Challenges of Human Relations in Education in the AI Era • The Normative Challenges of AI Literacy</i>
Systems: Tools and Technologies for Responsible AI	Open Track
Security: Re-evaluating Best Practices of Cybersecurity	<i>Let's Reevaluate Cybersecurity Best Practices</i>

Together, these chapters—*Ethics, Society, Education, Systems, and Security*—provide a coherent yet non-overlapping map for the papers of this volume.

The review effort behind ETHICOMP 2025 was remarkable. In the extended abstract stage, we handled 109 submissions, each of which was double-blind reviewed by at least two reviewers and then reassessed in a second round. Authors whose work matured through this process submitted 58 full papers; these underwent a separate evaluation, resulting in 51 papers in these LNCS proceedings and a parallel collection of approximately 90 extended abstracts accepted for presentation. The total number of reviews reached 422. A total of 167 submissions resulted in 51 LNCS published papers, yielding a global acceptance rate of approximately 30.5%, reflecting both peer review and authors’ availability.

Notably, authors were represented from 29 countries, and PC members came from 14 countries.

Acknowledgements

We thank all authors and reviewers, whose enthusiasm and careful work give *ETHICOMP* its substance.

Our thanks also go to the EasyChair team, whose platform handled an exceptional submission load, providing prompt technical support and error recovery, and ultimately helped produce the publication-ready materials for these volumes.

We gratefully acknowledge *Simon Rogerson*: his founding vision for *ETHICOMP* and his thoughtful comments have strengthened this preface.

We are grateful for the help and support received from the organizing institution:

- Universidade Autónoma de Lisboa, Portugal.

Besides our university, the conference received help from various collaborators from the following institutions:

- Universidad Complutense de Madrid, Spain
- Universitat Rovira i Virgili, Tarragona, Spain
- Universidad de Barcelona, Spain
- Instituto Politécnico de Tomar, Portugal
- Centro de Investigação em Cidades Inteligente, Tomar, Portugal
- Centro de Investigação Autónoma TechLab, Lisboa, Portugal
- Center for Computing and Social Responsibilities, DeMontfort University, Leicester, UK
- Instituto Politécnico do Porto, Portugal
- COMEGI – Universidade Lusíada de Lisboa, Portugal
- Meiji University, Tokyo, Japan
- Southern Utah University, USA

June 2025

Isabel Alvarez
Mario Arias-Oliva
Adrian-Horia Dediu
Nuno Silva

Organization

Organizing Committee Chairs

Reginaldo de Almeida	Universidade Autónoma de Lisboa, Portugal
Mário Marques da Silva	Universidade Autónoma de Lisboa, Portugal

Organizing Committee Members

Isabel Alvarez	Universidade Autónoma de Lisboa, Portugal
Mario Arias-Oliva	Complutense University of Madrid, Spain
Raquel Cabeças	Universidade Autónoma de Lisboa, Portugal
Adrian-Horia Dediu	Universidade Autónoma de Lisboa, Portugal
Cristina Dias	Universidade Autónoma de Lisboa, Portugal
António Duarte Santos	Universidade Autónoma de Lisboa, Portugal
Joana M. Matos	Universidade Autónoma de Lisboa, Portugal
Kiyoshi Murata	Meiji University, Japan
Pedro Santos	Universidade Autónoma de Lisboa, Portugal
Manuel Serejo	Universidade Autónoma de Lisboa, Portugal
Nuno Silva	Universidade Lusíada de Lisboa, Portugal
Ana Trindade	Universidade Autónoma de Lisboa, Portugal
Diogo Trindade	Universidade Autónoma de Lisboa, Portugal

Program Committee Chairs

Isabel Alvarez	Universidade Autónoma de Lisboa, Portugal
Mario Arias-Oliva	Complutense University of Madrid, Spain
Adrian-Horia Dediu	Universidade Autónoma de Lisboa, Portugal
Nuno Silva	Universidade Lusíada de Lisboa, Portugal

Steering Committee

Andrew Adams	Meiji University, Japan
Mario Arias-Oliva	Complutense University of Madrid, Spain
Katleen Gabriels	Maastricht University, Netherlands
Shalini Kesar	Southern Utah University, USA

Kai Kimppa	University of Turku, Finland
Wilhelm E. J. Klein	Zetamotion, UK
Ana Maria Lara	University of Burgos, Spain
Kiyoshi Murata	Meiji University, Japan
Erica L. Neely	Ohio Northern University, USA
Jorge Pelegrín-Borondo	La Rioja University, Spain
Nuno Silva	Universidade Lusíada de Lisboa, Portugal
Sabina Szymoniak	Częstochowa University of Technology, Poland
Richard Volkman	Southern Connecticut State University, USA

Program Committee

Ryoko Asai	Ruhr-Universität Bochum, Germany
Alicia Blanco	Rey Juan Carlos University, Spain
Miguel Camacho Ruiz	Complutense University of Madrid, Spain
Ramón A. Carrasco	Complutense University of Madrid, Spain
Alejandro Cataldo	University of Talca, Chile
Pedro Cuesta Valiño	University of Alcalá de Henares, Spain
Jorge de Andrés Sánchez	Universitat Rovira i Virgili, Tarragona, Spain
Maria del Pilar Martínez Ruiz	University of Castilla-La Mancha, Spain
Ana Isabel Dórico Parada	University of Vigo, Spain
Antonio Fernández Portillo	University of Extremadura, Spain
William Fleischman	Villanova University, USA
Yasunori Fukuta	Meiji University, Japan
Katleen Gabriels	Maastricht University, Netherlands
Don Gotternbarn	East Tennessee State University, USA
Pablo Gutiérrez Rodríguez	University of León, Spain
Frank Iorfino	University of Sydney, Australia
Pedro Isidoro González Ramírez	Universidad Autónoma de San Luis Potosí, Mexico
Younes Karrouk	Université Abdelmalek Essaadi, Morocco
Shalini Kesar	Southern Utah University, USA
Jani Koskinen	University of Turku, Finland
Orlando Lima Rua	Polytechnic Institute of Porto, Portugal
Antonio Marturano	University of Rome Tor Vergata, Italy
Joana M. Matos	Universidade Autónoma de Lisboa, Portugal
Kiyoshi Murata	Meiji University, Japan
Luis Oliveira	Polytechnic Institute of Tomar, Portugal
Yohko Orito	Ehime University, Japan
Graciela Padilla Castillo	Complutense University of Madrid, Spain
Ugo Pagallo	University of Turin, Italy

Ludovica Paseri	University of Turin, Italy
Jorge Pelegrín-Borondo	University of La Rioja, Spain
Antonio Pérez-Portabella	Universitat Rovira i Virgili, Tarragona, Spain
Adam Poulsen	University of Sydney, Australia
Lourdes Rivero Gutiérrez	Rey Juan Carlos University, Spain
Lorenzo Rodríguez Comesaña	University of Vigo, Spain
Simon Rogerson	De Montfort University, UK
Ziwei Shu	Complutense University of Madrid, Spain
Mar Souto-Romero	Rey Juan Carlos University, Spain
Mamello Thinyane	University of South Australia, Australia
Luis Valadares Tavares	Universidade Lusíada de Lisboa, Portugal
Kirsten Wahlstrom	University of South Australia, Australia
Marty J. Wolf	Bemidji State University, USA
Juan Carlos Yañez Luna	Universidad Autónoma de San Luis Potosí, Mexico
Xenia Ziouvelou	National Center for Scientific Research Demokritos, Greece

Additional Reviewers

Ala Almahameed	Sze Hwee Jace Tay
Inês Almeida	Jose Javier Galan Hernandez
Oliver Arroyo	Alexios Kaponis
Eleonora Bassi	Michael Kirkpatrick
Jacopo Ciani	Mário Marques da Silva
Laercio Cruvinel Júnior	Parag Mehta
Bárbara da Rosa Lazarotto	Tara Miranović
Antonio Dante Maria Santangelo	Nikos Nikoletos
Kathy Darzanos	Héctor Orrillo
Massimo Durante	Michael Papademas
Nina Evans	Carlos Porcel
Gema Fernández-Avilés	Aleksandra Pyrkosz-Dziubczyk
Federico Galli	Antonio Segura
Adrian Gavornik	Jesus Serrano-Guerrero
Konstantina Giouvanopoulou	Olga Siedlecka-Lamch
Miguel Guillén Pujadas	William Steingartner
Hanane Haddad	Sabina Szymoniak
Olli I. Heimo	Teresa Torres
Sarah Hladikova	Jorge Velasco
Miguel Houghton	Alejandra Vera González

Contents

Ethics: AI Ethics, Governance, and Values

Ethical Principles for the Production of Official Statistics Using Machine Learning and Artificial Intelligence Techniques	3
<i>Jorge-Eusebio Velasco-López, Ramón-Alberto Carrasco, Gema Fernández-Avilés, and José-María Montero-Lorenzo</i>	
Identifying AI Challenges in Research Practices Through Research Ethics Reviews	15
<i>Konstantina Giouvanopoulou, Xenia Ziouvelou, Michael Papademas, Kostas Karpouzis, and Vangelis Karkaletsis</i>	
Interpretability and the Measurement of Ethical Foundations in Artificial Intelligence	27
<i>Miguel Houghton Torralba, Ziwei Shu, Ramón-Alberto Carrasco, and María Francisca Blasco López</i>	
The Mediating Effect of Job Crafting in the Relationship Between Organizational Commitment and Organizational Citizenship Behavior and Its Ethical Implications	35
<i>Melissa Eugenia Treviño Serna, Rosalba Martínez Hernández, and Juan Carlos Yáñez Luna</i>	
On the Current (Im) possibility of Achieving Public Value Through the EU's Digital Strategy: An Ethics Method to Seek a "Collectual" Equilibrium	48
<i>Stefano Calzati</i>	
Inclusive Governance of Artificial Intelligence: Towards an Ethical Framework for Neurodivergence	60
<i>Elisabet Bolarín-Miró, María Mercedes Carmona Martínez, and Asunción López-Arranz</i>	
How AI is Reshaping Creativity: DeepSeek vs ChatGPT Plus in LEGO® SERIOUS PLAY®	69
<i>António Almeida, Denise Meyerson, and Inês Almeida</i>	

Sovereignty, Surveillance, and the Cloud: Geopolitical and Ethical Issues
of Global Cloud Computing 79
*Ala Ali Almahameed, Jorge Pelegrín-Borondo,
Juan Luis López-Galiacho Perona, and Mario Arias-Oliva*

Society: Social Networks, Youth, and Gender in our Digital Society

The Factor of Social Media and Artificial Intelligence in the Radicalization
of Young People in Jihadist Terrorism 95
Paula M. Núñez-Guerra

Using AI for Research and Educational Support Enhancing the Design
of Computer-Based Evaluations 105
Joana M. Matos and Adrian-Horia Dediu

Towards a Gender-Inclusive Tech Landscape in Portugal: Women4Digital’s
Insights on Gender and Digital Transformation 116
Rosa Monteiro, Mariana Santos, and Lina Coelho

Integrating Ethics and Gender Equality in Artificial Intelligence Education:
A Study of Higher Education in Portugal 129
Camila Marques, Rosa Monteiro, and Nuno Lourenço

Gender and Emerging Digital Technologies in Education 141
Davorka Radaković and William Steingartner

Impact of Gender Bias in the Output of AI Language Models on Heavy
Users 153
Sarah Hosell

Education: Learning and Literacy in the AI Era

Students’ Perception of the Integration of GenAI in Academic Paper
Assignment Preparation 169
Katerina Zdravkova and Bojan Ilijoski

Comparative Analysis of Instructor and AI Assessments: Objectivity,
Biases, and Impact on Academic Grading 181
Teresa Torres-Coronas

AI Ethics in Higher Education: A Review of Ethical Challenges 192
*Laercio Cruvinel Júnior, Orrillo Héctor Ascama,
and Mário Marques da Silva*

Just Hallucinations? The Problem of AI Literacy with a New Digital Divide ...	204
<i>Eleonora Bassi and Ugo Pagallo</i>	
Ethical Aspects of Distributed Extended Reality Training	215
<i>Olli I. Heimo and Teijo Lehtonen</i>	
Corporate Financial Statement Analysis in Education 4.0	227
<i>Miguel Guillén-Pujadas, David Alaminos, Álvaro Carrasco-Aguilar, and Mar Souto-Romero</i>	
Challenging AI as Critical Thinking	239
<i>Michael S. Kirkpatrick</i>	
AI Ethics in Higher Education Content Creation	252
<i>Álvaro Carrasco-Aguilar, Miguel Camacho-Ruiz, María Concepción Parra Meroño, and María Mercedes Carmona Martínez</i>	
GenAI and Proportionality: European and Portuguese Ethical-Legal Framework	263
<i>Filipa Pais D'Aguiar, Fernanda Duarte, and Nuno Silva</i>	
Rethinking Educational Assessment in the Age of Artificial Intelligence	274
<i>Mario Arias-Oliva, Antonio Pérez-Portabella, Manuel Ollé Sesé, and Jorge de Andrés Sánchez</i>	
Systems: Tools and Technologies for Responsible AI	
Evaluating the Role of Chatbots in Higher Education Based on Students' Experiences	289
<i>Ewa Duda, Ewelina Młynarczyk-Karabin, and Julia Więckiewicz-Modrzejewska</i>	
Parental Memory and Digital Traces of School Closures During the COVID-19 Pandemic: What Is Remembered, What Fades, and What Is Left Behind	301
<i>Akiko Orita</i>	
“There Are so Many” - Harms of Smart Homes	314
<i>Sally Bagheri</i>	
Digital Identity and Control: How AI Replicas Challenge Performance Rights	326
<i>Ahmed Qayyum and Stacy A. Doore</i>	

Through the Educators' Lens: University Teachers' Perceptions of AI Integration in Higher Education	338
<i>Ewa Duda, Ewelina Młynarczyk-Karabin, and Julia Więckiewicz-Modrzewska</i>	
Ethical Issues in the Use of Generative AI Chatbots for Therapeutic Purposes	350
<i>Anne Gerdes</i>	
A Case Against the Feasibility of AI Consciousness (AIC)	360
<i>Bo Kampmann Walther and Anne Gerdes</i>	
Some Problems in the Ethical Impact Assessment of Emerging Technologies and Socio-Technical Visions: Case CityVerse	371
<i>Antero Karvonen, Jaana Leikas, Nina Wessberg, Anton Sigfrids, and Sanni Pöntinen</i>	
Security and Ethics in the Use of Computing Technologies and the Internet	385
<i>Laercio Cruvinel Júnior, António Cabeças, and Adriana Lopes Fernandes</i>	
Social Risks of Brain-Machine Interface Usage: Questionnaire Survey of People with and without Disabilities	397
<i>Yohko Orito, Tomonori Yamamoto, Shizuka Suzuki, Hidenobu Sai, Kiyoshi Murata, Yasunori Fukuta, Taichi Isobe, and Masashi Hori</i>	
Sustainable Fashion Trends on Second-Hand Shopping: The Role of Social Media	412
<i>Lídia Monteiro, Orlando Lima Rua, Mar Souto-Romero, and Rafael Canorea-García</i>	
The Nexus Between Artificial Chatbots and Individual Creativity	420
<i>Luana Coelho, Orlando Lima Rua, and Felipe Debasa</i>	
Symbolic Aspects of Online Privacy Protection Behaviour: From a Social Communication Perspective	428
<i>Yasunori Fukuta, Kiyoshi Murata, Yohko Orito, Vanessa Bracamonte, and Takamasa Isohara</i>	
Operational Archives and the Right to Be Forgotten	441
<i>Klara Källström</i>	
Smart Doorbells in a Surveillance Society	451
<i>Anjela Mikhaylova, Sara Wilford, Bernd Stahl, and Laurence Brooks</i>	

User Engagement and Barriers in the Standardization Processes of Digital ID Architectures	464
<i>Yoshiaki Fukami and Kazue Sako</i>	
Systematic Review on AI Ethics in Privacy for V2X Communication	476
<i>Héctor Dave Orrillo Ascama, Laercio Cruvinel Júnior, and Mário Marques da Silva</i>	
Beyond Regulation and Moderation: A Forster-Inspired Framework for Machine Evolution	490
<i>William Fleischman and Leah Namisa Rosenbloom</i>	
The Epistemic Politics of Biometric Border Control	504
<i>Marie Eneman and Jan Ljungberg</i>	
Quizly: Transforming Quiz Experiences with Multi Modal Inputs for Differently Abled Users	516
<i>Farzana Jabeen, Sidra Sultana, Tahira Anwar Lashari, and Mehvish Rashid</i>	
Security: Re-evaluating Best Practices of Cybersecurity	
Comprehensive Approaches to Personal Data Protection Amid Evolving Cyber Threats	531
<i>Sabina Szymoniak and Mariusz Kubanek</i>	
Deepfake Manipulation and Ethical Dilemmas: A Comprehensive Risk Assessment	542
<i>Mariusz Kubanek and Sabina Szymoniak</i>	
Digital Agriculture Under Threat: Cybersecurity Challenges and Policy Gaps	554
<i>Olga Siedlecka-Lamch</i>	
Cybersecurity 2030: The Synergy Between Machine Learning and Generative Artificial Intelligence	566
<i>Mário Marques da Silva, Fernando Correia, Héctor Dave Orrillo Ascama, Laercio Cruvinel Júnior, Guilherme Lopes Fernandes, Tomás Viana, and Afonso Frasilho</i>	
Artificial Intelligence and Ethical Responsibility: The Impact of Algorithmic Decisions on Cybersecurity	579
<i>Sabina Szymoniak and Shalini Kesar</i>	

Cybersecurity, Best Practices: A Comprehensive Guide 589
 Pedro Brandao

Enhancing Health Information Access via ChatGPT and the E-Citizens
Portal 599
 Natalija Zeko, Darko Galinec, and William Steingartner

Author Index 599

Ethics: AI Ethics, Governance, and Values



Ethical Principles for the Production of Official Statistics Using Machine Learning and Artificial Intelligence Techniques

Jorge-Eusebio Velasco-López¹(✉), Ramón-Alberto Carrasco²,
Gema Fernández-Avilés³, and José-María Montero-Lorenzo³

¹ Faculty of Statistics, Universidad Complutense de Madrid, Madrid, Spain
jorge.velasco.lopez@ine.es

² Department of Marketing, Faculty of Statistics, Universidad Complutense
de Madrid, Madrid, Spain
ramoncar@ucm.es

³ Department of Applied Economics I, Faculty of Law and Social Sciences, University of
Castilla-La Mancha, Toledo, Spain
{Gema.FAviles, Jose.MLorenzo}@uclm.es

Abstract. The development and implementation of Artificial Intelligence (AI) and Machine Learning (ML) systems in the field of official statistics requires an ethical and regulatory approach to ensure the protection of human rights, privacy, and fairness. The 2024 Artificial Intelligence Strategy (Gobierno de España, 2024) emphasizes the importance of responsible AI in guaranteeing that these systems are safe, transparent, and beneficial to society. Given the potential impact of AI, organizations must establish an ethical framework to manage the risks associated with its use.

The paper proposes an ethical framework based on international and European principles for AI, such as transparency, impartiality, and human oversight. It also addresses the challenges of AI adoption, including technical competence, infrastructure, and governance. The goal is to ensure that AI-ML systems enhance the quality and efficiency of the production of official statistics, without compromising public trust or ethical standards.

Keywords: Ethical principles · Official Statistics · Artificial Intelligence

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have rapidly become transformative forces across various fields and industries, including official statistics. The ability of AI-ML systems to process large datasets, identify patterns, and generate predictive insights has opened up new possibilities for statistical agencies. These technologies offer the potential to enhance efficiency, accuracy, and relevance in statistical production, enabling data-driven decision-making at unprecedented scales.

However, the integration of AI-ML in official statistics also presents significant challenges relating to ethical considerations, data governance, and the need to maintain public trust, among other issues.

The primary objective of this paper is to examine the possible implementation of an AI policy in the context of official statistics, highlighting the elements that an AI policy should include, such as key life cycle phases such as planning, design, development, and implementation. It also explores the ethical principles that must be observed to ensure the responsible use of AI-ML technologies in statistical processes. Given that official statistics serve as the foundation for policy-making, economic planning, and social research, it is crucial to maintain high standards of reliability, transparency, and fairness.

The ethical dimension of AI-ML adoption is particularly relevant as statistical agencies must balance technological innovation with principles such as data privacy, accountability, and proportionality.

The paper proposes a framework of ethical principles tailored to the statistical domain, aimed at ensuring that AI-ML applications align with existing statistical standards and best practices. By adhering to these principles, national statistical offices (NSOs) can harness the power of AI while upholding their commitment to producing objective, impartial, and high-quality statistics.

At the international level, various institutions have developed ethical principles for the use of AI in statistics. For example, the United Nations Educational, Scientific, and Cultural Organization (UNESCO), with its Recommendation on the Ethics of Artificial Intelligence [1], and the Organisation for Economic Co-operation and Development (OECD), with its AI Principles [2], have outlined guidelines for the ethical development and application of these systems. In the European Union (EU), the High-Level Expert Group on AI has defined guidelines for trustworthy AI based on the principles of transparency, fairness, and human oversight [3].

Ethical models proposed by international organizations agree on the necessity of establishing a robust governance framework for AI in official statistics. This framework should include mechanisms for human oversight, transparency in algorithmic processes, respect for data privacy, and continuous evaluation of AI's social impacts. Additionally, it should promote inclusivity and diversity in AI systems, preventing discriminatory biases and ensuring equitable access to information.

The 2019 Ethics Guidelines for Trustworthy AI aim to promote trustworthy AI by identifying three essential components of trustworthiness throughout the system's life cycle: (a) AI must be lawful, complying with all applicable laws and regulations; (b) it must be ethical, respecting ethical principles and values; and (c) it must be robust, both technically and socially, as even AI systems designed with the best of intentions can cause unintended harm.

The guidelines also outline seven non-binding ethical principles that AI systems should adhere to in order to be deemed trustworthy. Many of these principles are already reflected in existing legal requirements, thus falling under the category of "lawful AI," the first component of trustworthy AI. However, while numerous legal obligations reflect ethical principles, adherence to these principles extends beyond mere legal compliance.

At the national level, Spain has integrated ethical principles into the production of official statistics, in line with the European Statistics Code of Practice (ESCOp) [4] and the Regulation (EC) No. 223/2009 [5]. These principles include professional independence, impartiality, objectivity, reliability, statistical confidentiality, transparency, and proportionality. The application of ML-AI to official statistics must adhere to these principles to ensure the quality and reliability of the information produced.

To develop the ethical framework proposed in this paper, a qualitative review was conducted of key international guidelines on AI ethics. Documents were selected based on institutional authority, policy relevance to official statistics, and alignment with data governance principles. The comparison of guidelines focused on ethical dimensions such as fairness, transparency, privacy, and human oversight.

2 Challenges of Artificial Intelligence for Official Statistics

The adoption of AI in official statistics presents significant opportunities but also poses several challenges that must be carefully addressed [6]. While AI has the potential to enhance data collection, processing, analysis, and dissemination, the risks can be substantial. To successfully integrate AI into official statistics, NSOs must address the following key challenges.

- **Building AI Competencies:** NSOs need to develop skilled workforces with technical, analytical, and ethical expertise in AI. Training should focus on data science, ML, visualization, and AI ethics. Continuous professional development and AI literacy programs are vital.
- **Infrastructure Development:** Many NSOs, especially in developing regions, lack access to adequate infrastructure essential for AI, such as electricity, internet, and computing resources. Investments in internet access, cloud computing, and data processing capacity are necessary to bridge the digital divide.
- **Compliance and Ethical Use:** The application of AI in statistics must adhere to legal and ethical standards such as General Data Protection Regulation (GDPR). Policies should ensure data confidentiality, transparency, fairness, and avoid bias. NSOs should embed ethical AI principles and strong data protection protocols.
- **AI Governance:** Effective AI adoption requires structured governance including AI steering committees, defined roles (e.g., AI officers), cross-departmental collaboration, and robust data governance policies. Global frameworks aligned with the UN Sustainable Development Goals (SDGs) are recommended.
- **Life Cycle Management of AI Projects (CDAC):** The integration of AI and ML in official statistics must follow a structured life cycle to ensure the responsible and ethical use of these technologies. The CDAC AI Life Cycle [7] is structured into three primary phases: design, develop, and deploy. Roles include AI/data scientists, ML engineers, legal/ethics advisors, and project managers. Each stage must be well-documented and professionally executed to ensure successful implementation.
- **Other Key Elements of AI Policy:** Technical standards for validated AI methodologies and risk mitigation protocols; controls and limits established to ensure human oversight, clear AI application boundaries, and mechanisms for handling errors or biases; communication and awareness-raising to engage internal staff, stakeholders, and the public in order to promote understanding, trust, and transparency in AI use.

By integrating these technical, governance, and communication elements, NSOs can ensure AI adoption enhances statistical accuracy, efficiency, and public trust while mitigating risks and maintaining ethical standards.

3 Proposed Ethical Framework for AI and ML in Official Statistics

This section reviews the commonly accepted ethical principles for responsible AI, followed by an overview of AI and Large Language Models (LLMs) principles compared with principles and international guidelines for official statistics. The section ends with a proposal for integrating AI and LLM considerations into ESCoP.

Although this paper primarily draws from existing regulatory and institutional frameworks, many of the ethical principles align with classic moral theories. For example, the emphasis on ‘do no harm’ and human dignity reflects deontological ethics, while the principle of fairness can be associated with utilitarian or Rawlsian perspectives. Recognizing these underpinnings enriches the conceptual grounding of the proposed ethical model.

The operationalization of ethical AI principles in official statistics can be better understood through practical examples. For instance, Statistics Netherlands (CBS) has implemented the use of ML for fraud detection in business surveys while maintaining transparency by publishing model documentation and bias mitigation protocols. Similarly, Statistics Canada has piloted AI-based imputation methods in censuses, guided by internal ethical review boards. These experiences demonstrate the feasibility of integrating responsible AI practices into core statistical workflows and highlight the role of governance mechanisms in ensuring ethical compliance.

3.1 Commonly Accepted Ethical Principles for Responsible AI

Ethics Guidelines for Trustworthy AI (High-Level Expert Group on Artificial Intelligence, EU)

The EU’s Ethics Guidelines for Trustworthy AI, published in 2019, aim to ensure that AI systems are lawful, ethical, and robust throughout their life cycle. Lawfulness requires compliance with existing regulations, ethicality demands respect for human rights and values, and robustness means systems must be technically sound and socially reliable. These three elements must work together, and any conflicts between them should be resolved through balanced solutions.

Grounded in the EU Charter of Fundamental Rights, the guidelines emphasize key principles: respect for human autonomy, prevention of harm, fairness, and explicability. While many are tied to legal obligations, the guidelines underscore that ethical AI goes beyond the law, stressing the importance of focusing on doing what is inherently right even if regulations are not yet in place.

AI Regulation – Recital 27 of the EU AI Act [8]

Recital 27 of the EU AI Act outlines seven ethical principles to ensure trustworthy, human-centered AI aligned with EU values. Specifically, AI systems must support human

agency; be technically robust and safe; respect privacy and data governance; be transparent and explainable; promote fairness and non-discrimination; contribute to social and environmental well-being; and ensure accountability. These principles guide responsible AI design, governance, and regulation across the EU.

UNESCO's Recommendation on the Ethics of Artificial Intelligence [1]

UNESCO's 2021 Recommendation on the Ethics of AI offers a global framework to guide responsible AI use. It promotes ethical AI through the Global AI Ethics and Governance Observatory, helping countries assess and improve AI governance.

The framework is grounded in four core values: respect for human rights and dignity, promotion of just societies, diversity and inclusion, and environmental sustainability. It also sets out 10 guiding principles; namely, do no harm, safety, privacy, inclusive governance, accountability, transparency, human oversight, sustainability, public awareness, and non-discrimination. These principles are aimed at ensuring AI development benefits humanity, protects rights, and fosters global equity.

OECD AI Principles [2]

The OECD AI Principles, updated in 2024, offer a global framework for ethical AI. They emphasize inclusive growth, respect for human rights, transparency, robustness, and accountability. These principles guide the responsible development and use of AI across sectors.

To support adherence to these principles, the OECD recommends investing in AI R&D, fostering inclusive ecosystems, aligning international governance, building workforce skills, and enhancing global cooperation. The goal is to ensure AI is fair, transparent, and beneficial to all.

Asilomar AI Principles (2017) [9]

The Asilomar AI Principles, established in 2017, outline 23 guidelines to ensure AI is developed responsibly, safely, and in alignment with human values. They are organized around three core themes. In the area of AI research, the principles call for development to prioritize human well-being, support responsible investment that includes ethical and societal considerations, foster collaboration and transparency, encourage engagement between researchers and policymakers, and work to prevent competitive dynamics that could lead to unsafe systems. On the ethical and value-based front, the principles emphasize the importance of safety, reliability, and transparency; advocate for AI that supports human dignity and cultural diversity; and call for fair distribution of AI's benefits. They also stress the need for human oversight and the prevention of misuse.

Finally, the long-term aspects of the framework aim to ensure AI remains beneficial over time by addressing emerging risks before they become harmful. Together, these principles seek to guide AI development in a way that maximizes benefits while safeguarding human values and societal well-being.

3.2 Overview of AI and LLM Principles Compared with Principles and International Guidelines for Official Statistics

Applying these principles to official statistics requires the development of specific codes of conduct for the use of ML-AI, aligned with principles and international guidelines for

official statistics. These codes should establish clear standards for transparency, fairness, and data protection, ensuring that AI applications in official statistics are used responsibly and benefit society.

This proposed ethical framework for AI and ML is based on the statistical principles established in European and national legislation. The production of official statistics is governed by Article 338.1 of the Treaty on the Functioning of the European Union, Regulation 223/2009 [5], and ESCoP [4], ensuring adherence to fundamental principles such as independence, impartiality, reliability, transparency, and proportionality.

Based on the analysis conducted, it is proposed that the ethical approach to be adopted by NSOs in their use of AI-ML should be developed from existing statistical principles. (The principle of cooperation has not been considered, as it is primarily relevant at the international level; at the national level, this obligation is already enshrined in law, along with the mechanisms for enforcing it.)

The production of official statistics is subject to well-established principles, which have been constitutionalized in Article 338.1 of the Treaty on the Functioning of the European Union, defined in Regulation (EC) No 223/2009 [5] on European statistics, and further elaborated in ESCoP [4]. These principles also take into account the Fundamental Principles of Official Statistics [10], adopted by the United Nations Economic Commission for Europe (UNECE) on April 15, 1992, and by the United Nations Statistical Commission on April 14, 1994. Therefore, national statistics must adhere to the principles of professional independence, impartiality, objectivity, reliability, statistical confidentiality, cost-effectiveness, transparency, specialization, and proportionality. In addition to these principles, ESCoP incorporates the quality criteria established in Article 12 of Regulation 223/2009 [5]: relevance, accuracy, timeliness, punctuality, accessibility and clarity, comparability, and coherence.

The Quality Assurance Framework (QAF) of the European Statistical System (ESS) [11], is a set of guidelines designed to help ensure the quality of statistics produced within the ESS. The ESS includes NSOs from EU member states, as well as other statistical bodies working under the auspices of Eurostat, the statistical office of the EU.

The use of AI-ML systems has sparked numerous debates and proposals regarding the implementation of ethical codes and standards by organizations that develop and utilize these technologies. Many of the principles established in these codes—and recognized by international organizations such as the United Nations (UN), the OECD, and even the EU—share common ground.

However, in the statistical domain, these principles must be analyzed and interpreted in the context of existing statistical principles. Therefore, when using AI-ML for the design, production, and dissemination of official statistics, NSOs will have to adapt some of these ethical principles. Table 1 presents a comparison of the shared principles established under UNECE LLM Guidelines [12] and those outlined in ESCoP [4].

In terms of similarities between the two frameworks, many principles overlap, especially in regard to fairness, privacy, proportionality, transparency, and accountability. The main difference between the two is that ESCoP focuses on statistical methodologies and

Table 1. Comparison of Principles: UNECE LLM Guidelines vs. ESCoP

UNECE LLM Guidelines	European Statistics Code of Practice (ESCoP)	Similarities	Differences and Adaptations
Do no harm: LLMs should not cause intended or unintended harm, including bias. Projects should define expected benefits and establish risk mitigation	Statistical Confidentiality and Data Protection: Ensures privacy, confidentiality, and security of data providers and prohibits misuse of statistical data	Both frameworks emphasize the prevention of harm and the guarantee of data security	UNECE explicitly addresses AI-related risks, while ESCoP focuses on data security and confidentiality in statistics
Defined purpose, necessity, and proportionality: Use of LLMs must be proportionate to needs and should not undermine human authority or rights	Proportionality: Information collected must balance necessity and intended outcomes	Both stress the importance of proportionality in data use and decision-making	ESCoP applies proportionality specifically to data collection, while UNECE extends it to AI's impact on human rights
Safety and security: All projects must consider cybersecurity and information security risks, including potential misuse of LLMs	Statistical Confidentiality and Data Protection: Guarantees the protection of data from unauthorized use	Both stress the importance of security in data handling	UNECE extends security considerations to cybersecurity risks in AI systems
Fairness and non-discrimination: LLMs are prone to biases; these must be evaluated and mitigated to ensure fairness	Impartiality and Objectivity: Statistics must be compiled neutrally and treat all users equitably	Both frameworks require neutrality and the mitigation of biases	UNECE focuses on algorithmic biases in AI, whereas ESCoP addresses fairness in statistical reporting
Sustainability: AI use should consider environmental and social impacts, including energy consumption and long-term consequences	Cost-Effectiveness: Statistical resources should be used efficiently to minimize costs	Both focus on efficient resource use	ESCoP focuses on economic efficiency, while UNECE extends to environmental and social sustainability

(continued)

Table 1. *(continued)*

UNECE LLM Guidelines	European Statistics Code of Practice (ESCoP)	Similarities	Differences and Adaptations
Right to privacy, data protection, and governance: Privacy and data governance apply to LLMs’ training data and user inputs. Sensitive information must be protected	Statistical Confidentiality and Data Protection: Ensures the privacy and security of collected statistical data	Both prioritize privacy and data governance	ESCoP is specific to statistical data, while UNECE covers AI training and user interactions
Human autonomy and oversight: AI systems must not have autonomous decision-making power; humans must oversee all AI outputs	Professional Independence: Ensures statistical authorities operate independently of political or private influence	Both frameworks emphasize human control over processes	ESCoP applies independence to statistical production, whereas UNECE applies it to AI decision-making oversight
Transparency and explainability: Users must understand LLM-generated material, including potential biases and limitations	Transparency: Data providers have the right to obtain complete information regarding data protection and purpose	Both require transparency in processes and outputs	UNECE applies this principle to AI explainability, while ESCoP focuses on transparency in data collection and use
Responsibility and accountability: AI use should be governed by clear accountability structures, ensuring that humans remain responsible for decisions	Commitment to Quality: Statistical authorities continuously improve processes and take responsibility for quality	Both emphasize responsibility and governance structures	UNECE applies accountability to AI oversight, whereas ESCoP relates it to statistical quality control
Inclusion and participation: Stakeholders must be consulted regarding LLM use, particularly in decision-making	Relevance: European statistics must meet the needs of users	Both frameworks emphasize responsiveness to stakeholder needs	ESCoP ensures statistical relevance, while UNECE highlights inclusive participation in AI processes

data collection processes, while UNECE emphasizes AI ethics, sustainability, and cybersecurity. Thus, some adaptations are deemed necessary: ESCoP could incorporate AI-specific risks, while UNECE might benefit from statistical methodology considerations to ensure AI-generated insights align with rigorous data standards.

3.3 Proposal for Integrating AI and LLM Considerations into the European Statistics Code of Practice (ESCoP)

As AI and LLMs become increasingly relevant in the field of statistics, it is essential to ensure their responsible integration into the ESS. While AI presents opportunities for improving efficiency and analysis, it must align with the core principles of statistical integrity, transparency, and public trust. This proposal suggests specific adaptations to ESCoP to account for AI-related risks, ethical and sustainability considerations.

Strengthening AI Governance in European Statistics: Professional Independence and Human Oversight

Current Principle: *PRINCIPLE 1 – PROFESSIONAL INDEPENDENCE. The professional independence of statistical authorities from other policy, regulatory or administrative departments and bodies, as well as from private-sector operators, ensures the credibility of European statistics.*

The Proposed Adaptation is to reinforce the **principle of professional independence**, ensuring that AI remains a tool rather than an authority in statistical decision-making:

- **Maintaining Human Oversight:** AI-generated statistical outputs must never replace expert judgment. Qualified statisticians should review, validate, and refine AI-generated data before they are published.
- **Restricting AI Decision-Making:** While AI can assist in statistical modeling and analysis, it should not autonomously influence methodologies, interpretations, or policy recommendations.

Addressing AI Risks and Ethical Considerations: Impartiality, Fairness and Non-Discrimination; and Transparency and Explainability

Current ESCoP Principle (1): *PRINCIPLE 6 – IMPARTIALITY AND OBJECTIVITY. Statistical authorities must produce and disseminate European statistics respecting scientific independence and in an objective, professional and transparent manner in which all users are treated equitably.*

The Proposed Adaptation is taking these measures to align with the **principle of impartiality and objectivity**, ensuring that AI-assisted statistical processes maintain fairness and credibility:

- **Regular Bias Audits:** AI-generated statistics should undergo systematic reviews to detect and correct biases inherent in training data or model outputs.
- **Inclusive AI Development:** To mitigate risks of linguistic and cultural bias, training datasets should be diversified, ensuring that AI models do not disproportionately favor English-language or Western-centric sources.

Current ESCoP Principle (2): *PRINCIPLE 15: ACCESSIBILITY AND CLARITY.* *European statistics should be presented in a clear and understandable form, disseminated in a suitable and convenient manner, and made available and accessible on an impartial basis with supporting metadata and guidance.*

The Proposed Adaptation is to reinforce **the principle of transparency**, these adaptations help prevent the misuse or misinterpretation of AI-generated statistics.

- **AI Output Disclosure:** Any statistics generated or assisted by AI should be explicitly labeled so users understand the role AI played in the process. Explainability. Statistical offices should provide clear documentation on how AI-generated insights were derived, including potential limitations such as bias, hallucinations, or incomplete training data.

Strengthening Data Protection and Security in AI-Driven Statistics: Statistical Confidentiality and Data Protection

Current ESCoP Principle: *PRINCIPLE 5 – STATISTICAL CONFIDENTIALITY.* *The privacy of data providers (households, enterprises, administrations and other respondents), the confidentiality of the information they provide and its use only for statistical purposes must be absolutely guaranteed.*

The Proposed Adaptation is to reinforce **the principle of statistical confidentiality**, ensuring that the adoption of AI does not compromise data privacy or security:

- **Robust Data Governance for AI:** AI tools used in statistical applications must strictly adhere to GDPR and anonymization protocols. Sensitive statistical data should never be used for unauthorized AI training.
- **Cybersecurity Protections:** Given the risks of AI exploitation, statistical authorities must implement advanced security measures to prevent unauthorized access, adversarial attacks, and data leaks in AI-driven processes.

Ensuring Sustainable and Responsible AI Use: Cost-Effectiveness and Sustainability

Current ESCoP Principle: *PRINCIPLE 10 – COST EFFECTIVENESS.* *Resources must be effectively used.*

The Proposed Adaptation is to extend **the principle of cost-effectiveness**, ensuring that AI adoption aligns with broader environmental and sustainability goals:

- **Environmental Impact Assessment:** When using AI in statistics, its carbon footprint should be considered, particularly regarding energy-intensive model training, storage, and computational processes.
- **Sustainable AI Infrastructure:** Statistical authorities should prioritize energy-efficient AI models and align data center operations with EU sustainability guidelines to minimize excessive energy and water usage.

Fostering Stakeholder Inclusion and Collaboration: Inclusion and Participation in AI-Driven Statistics

Current ESCoP Principle: *PRINCIPLE 11 – RELEVANCE. European statistics must meet the needs of users.*

The Proposed Adaptation is to strengthen **the principle of relevance**, ensuring that AI-driven statistics remain responsive to societal needs while fostering trust and inclusivity:

- **Stakeholder Engagement in AI Integration:** When incorporating AI into official statistics, it is crucial to consult policymakers, industry experts, and civil society organizations. AI-generated insights should complement—rather than replace—human expertise.
- **Public Awareness and AI Literacy:** NSOs should actively educate the public on how AI contributes to statistical processes, addressing concerns about misinformation, bias, and data reliability.

Integrating AI and LLMs into European statistics offers numerous advantages, but requires careful governance to align with ESCoP principles. By ensuring transparency, fairness, security, and sustainability, NSOs can harness AI's potential while maintaining public trust and compliance with European statistical standards. The proposed adaptations ensure that AI serves as a reliable tool, reinforcing rather than undermining the integrity of official statistics.

4 Conclusion

The integration of AI into official statistics offers transformative potential but is not without challenges. To fully harness AI's capabilities, NSOs must invest in human capital development, infrastructure, legal frameworks, and governance structures. By addressing these barriers systematically, AI can serve as a powerful tool for enhancing data quality, accessibility, and relevance in official statistics. A global collaborative effort among governments, international organizations, and private sector stakeholders is essential to ensure that AI adoption in official statistics is ethical, equitable, and aligned with the broader goals of sustainable development.

The use of AI-ML in official statistics must adhere to specific principles that ensure its ethical and responsible application. This ethical framework reinforces the commitment of public statistics to quality, reliability, and responsibility in the use of AI-ML in official statistics, ensuring alignment with national and international standards.

The adoption of AI-ML in official statistics represents a significant opportunity to enhance data analysis, improve efficiency, and support informed decision-making. However, these advancements must be implemented responsibly, ensuring adherence to ethical principles and statistical best practices. By aligning AI initiatives with the fundamental principles of official statistics, agencies can maintain public trust, safeguard data integrity, and harness AI's potential for the greater public good.

The successful integration of AI-ML requires a structured approach, encompassing robust governance, methodological rigor, and continuous oversight. Ethical considerations, such as privacy protection, transparency, and accountability, should remain at the

forefront of AI applications in statistical domains. As AI technologies evolve, statistical agencies must adapt, ensuring that their methodologies remain relevant while upholding the core values of official statistics.

In terms of limitations, this study focuses on normative frameworks and institutional guidelines without providing in-depth technical analysis of algorithmic processes or quantitative assessment of policy impacts. The proposed framework has yet to be empirically validated across diverse national contexts. Furthermore, sector-specific ethical concerns (e.g., environmental statistics, labor market indicators) were not individually examined, representing areas for future work.

Thus, future research could attempt an empirical validation of the proposed ethical framework in various national contexts. It could also explore the development of automated AI audit tools for statistical processes, and conduct studies of public perceptions of trust in AI-generated official statistics. Comparative studies of the ethics of AI implementation in EU and non-EU countries would also be valuable.

Ultimately, responsible AI adoption will enable statistical agencies to fulfill their mission of providing accurate, reliable, and high-quality statistical information. By embracing innovation while adhering to ethical guidelines, AI-ML can serve as a powerful tool in the advancement of official statistics, shaping evidence-based policy and decision-making for the future.

References

Although it is not an exhaustive list, below are references to ethical codes on the use of AI-ML and statistical principles.

1. UNESCO: Recommendation on the Ethics of Artificial Intelligence (2021)
2. OECD: OECD AI Principles (2019)
3. European Commission, “Ethics Guidelines for Trustworthy AI” (2019)
4. European Statistical System, “European Statistics Code of Practice” (2022)
5. European Union, “Regulation 223/2009 on European Statistics” 2009
6. Popoola, O.P.: Artificial Intelligence for Official Statistics: Opportunities, Practical Uses and, Challenges (2024)
7. De Silva, D., Alahakoon, D.: An artificial intelligence life cycle: From conception to production. *Patterns* **3**(6), 2666–3899 (2022)
8. EU Artificial Intelligence Act, “Recital 27 | EU Artificial Intelligence Act” (2024)
9. Asilomar, “AI Principles” (2017)
10. ONU: Fundamental Principles of Official Statistics (United Nations) (2013)
11. European Statistical System: Quality Assurance Framework of the European Statistical System (QAF) (2019)
12. UNECE_HLG-MOS: Large Language Models for Official Statistics (2023)
13. Gobierno de España: Estrategia de Inteligencia Artificial 2024 (2024)



Identifying AI Challenges in Research Practices Through Research Ethics Reviews

Konstantina Giouvanopoulou^{1,2(✉)}, Xenia Ziouvelou¹, Michael Papademas^{1,2},
Kostas Karpouzis², and Vangelis Karkaletsis¹

¹ Institute of Informatics and Telecommunications, National Centre for Scientific Research
Demokritos, Aghia Paraskevi, Greece

{kgiouvano, xeniaziouvelou, papademasmichael,
vangelis}@iit.demokritos.gr

² Department of Communication Media and Culture, Panteion University of Social and Political
Sciences, Kallithea, Greece
kkarpou@panteion.gr

Abstract. The rapid advancement of technology and the spread of Artificial Intelligence (AI) across various sectors of society are reshaping our daily lives and imply a series of emerging challenges. Given this fact, it is vital to detect and diagnose the nature of these challenges at an early stage, adopting a proactive approach to the design, development, and evaluation of AI systems; an ethics-by-design approach permeating all research phases. The challenges posed by the widespread use of AI systems impact the research ethics review process, which aims to ensure ethical and responsible research practices. Therefore, it is urgent to understand the effect of AI in the research area, along with the arising social implications, in order to increase the effectiveness of the research ethics review process for achieving more ethical AI research outcomes. This study aims to identify challenges related to the use of AI technologies across different sectors and to detect if and how these challenges impact research ethics and the research ethics review process. For this purpose, a systematic literature review was conducted. The initial findings of this review facilitated the creation of a classification framework, which provides a structured basis for in-depth research. The proposed framework is based on the assumption that identifying and categorising the arising challenges can enable more effective ways to address them. Thus, future policies to meet these challenges can be more targeted and impactful. This study highlights the need for further research, including a deeper and more extensive literature review in the field, as well as empirical validation of the classification framework through qualitative feedback from members of Research Ethics Committees (RECs).

Keywords: Research Ethics Reviews · Research Ethics Committees · AI-related challenges

1 Introduction

The swift advancement of Artificial Intelligence (AI) can undoubtedly lead to new opportunities enhancing our daily life, but the overuse of this technology can also result in potential risks and emerging challenges. On the other hand, widespread concerns about

its application may sometimes cause the under-utilisation of AI technologies [1]. The foregoing implies that there must be knowledge on ethical issues related to the use of AI, in order to identify potential challenges and risks in advance and to avoid extremes in its use, whether they concern overuse or underuse. Therefore, it is necessary to ensure that AI technology is always used towards ethical and socially beneficial outputs and that ethics permeate all stages of its development, from the design phase to the final deployment.

Over the last few years, as AI has been expanding rapidly affecting more and more aspects of our daily lives, many initiatives have emerged, from different actors, promoting the ethical use of AI and various guidelines and principles have been developed to ensure its ethical use and to address emerging ethical challenges [2–8]. With the aim of optimising AI research to develop socially beneficial applications, it is necessary to identify emerging challenges and mitigate the risks of AI systems. The arising challenges related to AI can be of different nature, such as technical, scientific, social, political, and the risks may be linked to the security of applications and the privacy of users. Researchers specialising in AI, on their side, should raise awareness regarding potential risks, and they ought to be engaged in integrated research to identify and address the social risks of AI, as the increasing effectiveness of AI is in proportion to its societal responsibility [9].

In light of the aforementioned, a particularly pressing and topical issue that arises is the review and evaluation of research proposals involving AI technologies as part of the research ethics review process conducted by the Research Ethics Committees (RECs). Research ethics relates to promoting good through research activities and outputs, aiming at positive social impact and at the protection of human rights; demonstrating respect for persons and their autonomy, safeguarding justice and beneficence (Belmont Report¹), and protecting interculturality and values. Research ethics should permeate the entire cycle of research activity, starting first and foremost with the researcher and including the chosen research methods, the research funding, and the role of RECs. The term “research ethics review” refers to the process of reviewing and evaluating the submitted research protocols with the aim of ensuring that the reported research is conducted in an ethical and responsible manner, protecting research participants, minimising the risk of their harm, and safeguarding their rights and dignity. Research ethics review processes assess the ethical aspects of submitted research to ensure compliance with the law, established ethical standards, and guidelines. In general terms, a standard research ethics review procedure involves the submission of the research protocol, along with all relevant and necessary documents, including information sheets and informed consent forms [10].

The oversight mechanisms responsible for the ethics review process are Research Ethics Committees (RECs), also known and referred to as Research Ethics Boards (REBs) or Institutional Review Boards (IRBs). They hold the pivotal role of ensuring that research is ethically designed and conducted according to regulations and ethical standards aimed at protecting all the participants [11] and avoiding research misconduct. RECs, by assuring compliance with commonly accepted rules and adherence to fundamental ethical principles and standards according to legal rules, soft law, and codes of

¹ The Belmont Report, *Ethical Principles and Guidelines for the Protection of Human Subjects of Research*.

conduct, ensure and safeguard the scientific validity of a project at the earliest stages, by providing a multidisciplinary approach to review [12]. RECs serve and perform their duties and tasks in research centres/institutions, governmental organisations, as well as within the private sector [13]. Their operation must be governed by effective procedures to support ethical research. In addition to their standard research ethics review procedures, RECs should be equipped to identify emerging ethical challenges specific to AI technologies. Their assessment processes must actively promote and protect responsible innovation.

2 Methodology

To identify the emerging ethical challenges that are associated with AI technology and how these challenges impact and are addressed through research ethics reviews, we conducted a systematic literature review (SLR) (Fig. 1) by specifying the following research questions to be examined:

- Which are the main challenges that RECs face during the research ethics review process, in the light of emerging technologies?
- Which are the emerging ethical research challenges of AI technology systems across different domains?
- And most importantly, how do AI and emerging technologies impact research ethics and research ethics review processes?

The SLR carried out was based on the guidelines for conducting systematic literature reviews by Kitchenham and Charters [14] and also followed PRISMA guidelines [15]. For the purposes of this research, the Scopus database was selected. The first search for articles was conducted on 19 August 2024 using the keywords: AI AND ethical AND challenges AND research AND ethics AND reviews (search within: Title, Abstract, Keywords). The query yielded 169 articles and after excluding any other language than English the query resulted in 162 documents. On 19 August, a *title detection* was carried out, and of the 162 articles resulting from the search, 66 were selected. The inclusion criteria were the following:

- titles referring to challenges, implications, biases, and ethics related to the use of AI in various fields,
- titles referring to research ethics, research ethics reviews, and research ethics committees.

Subsequently, an *abstract detection* was carried out on 20 August, which resulted in 56 articles out of 66. The selected inclusion criteria were limited to abstracts referring to challenges, implications arising from the application of AI technology in diverse fields, and challenges concerning research ethics and research ethics reviews. The *full text* detection was conducted from 21 August to 1 November 2024, and resulted in 30 documents according to the following inclusion and exclusion criteria:

Inclusion criteria:

- articles addressing the challenges, implications, and ethical issues of AI applications in various fields from a general perspective,

- articles concerning research ethics and research ethics review challenges.

Exclusion criteria:

- articles referring to challenges and implications of specific subcategories of the diverse fields were not included (articles referring to specific areas of a broader theme were excluded from the full text detection).

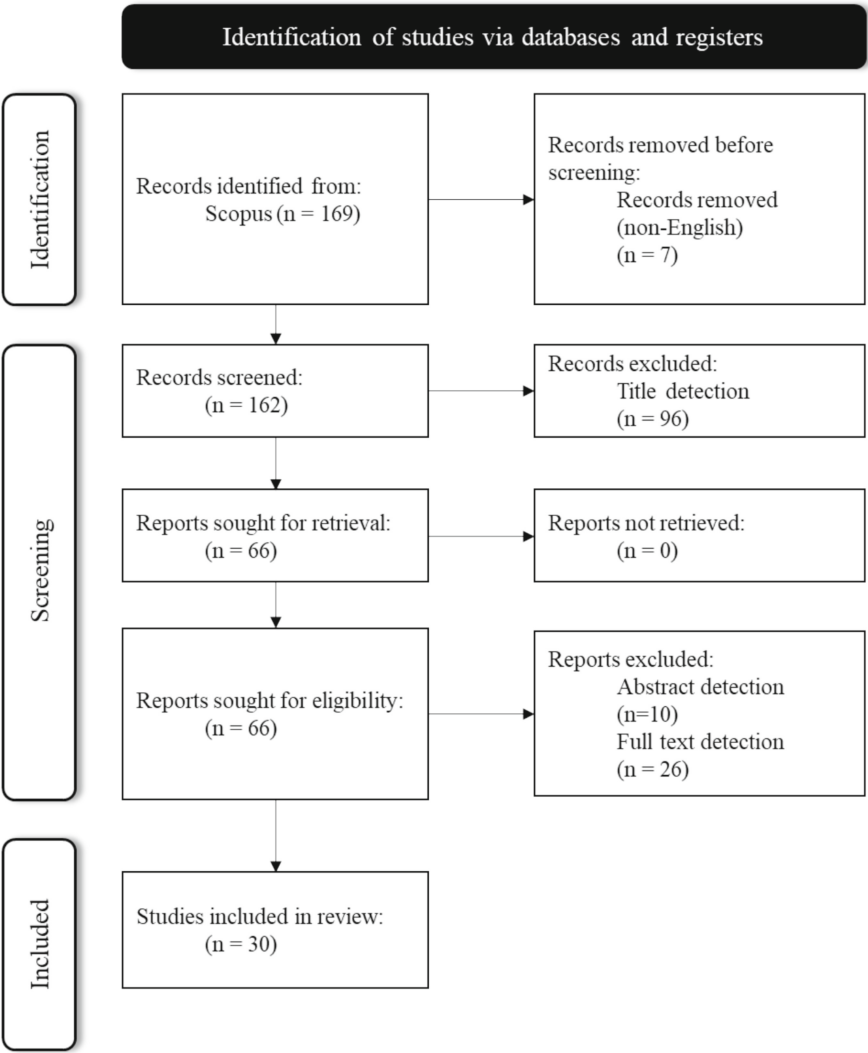


Fig. 1. Systematic Literature Review, PRISMA flowchart [15]

The articles that emerged from the systematic literature review provide important input on the arising challenges posed by the application of AI technology in different fields, and their categorisation allows their study by sector.

In addition to the systematic literature review that was conducted in the Scopus database, some articles were identified from the literature that are particularly influential and contribute to a more thorough study of the issue [16]. Accordingly, specific articles were selected to provide substantial information that complements our existing analysis, particularly with regard to general challenges associated with the composition of RECs and the procedures they follow.

3 Key Findings and Classification Framework

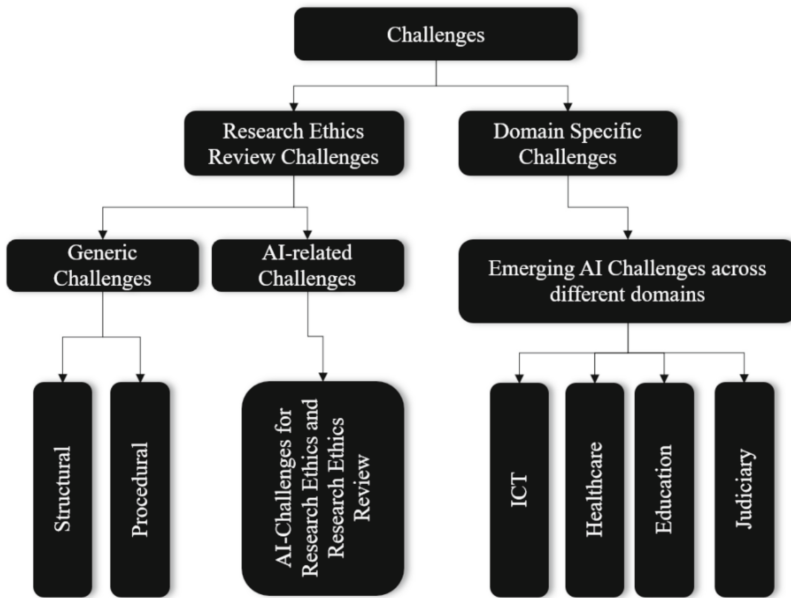


Fig. 2. Research Ethics Review and AI Challenges: A Classification Framework

The literature review enabled the identification of diverse levels of examination and informed the development of a Classification Framework (Fig. 2), which outlines the following types of challenges:

- **Research Ethics Review Challenges:** including generic challenges that are associated with the research ethics review, such as structural challenges that concern the composition and the expertise of the members forming the RECs, and procedural challenges focusing on the procedures followed by RECs for the evaluation of research proposals; and AI-related challenges, associated to research ethics that need to be identified and addressed through the research ethics review process.

- **Domain Specific Challenges:** including emerging AI challenges across different domains, arising from the use of AI technology in different fields (Information and Communication Technology (ICT), Healthcare, Education, Judiciary).

3.1 Research Ethics Review Challenges

Challenges related to the research ethics review may refer to some general challenges linked to RECs and challenges/implications that arise during the review process. In particular, the **structural** composition of RECs and the need for **multidisciplinarity** and **diverse expertise** is a requirement for their harmonious functioning and a more effective evaluation of proposals. It is essential to include members with diverse backgrounds in RECs [13] and the diversity of expertise is also a necessity in research involving the use of AI technologies [17]. This offers the potential to cover any lack of knowledge/training in research ethics, possibly observed in research involving AI technologies. The multidisciplinary nature of the structure of the RECs is particularly beneficial for a deeper reflection on ethical issues and considerations. The Research Ethics Committees, which are composed of a group of experts, are charged with the responsibility of diagnosing and reviewing ethical issues and addressing them from the outset, at the research design stage. Therefore, cooperation between their members and interdisciplinarity can more effectively anticipate imminent risks.

In addition, there are issues that affect the operation of RECs and lead to **procedural** challenges. **Bureaucratic issues** and the standardisation of forms, which may not be suitable for all types of research, can be an obstacle to the ethical review process, causing an administrative workload for researchers; other procedural issues may relate to delays due to slow RECs operations [10]. Procedural challenges, as identified in our classification, may arise from either **weak regulation** or, conversely, from **over-regulation** resulting in overlapping and burdensome regulation even for non-high-risk areas of research [18]. **Inconsistencies in decision-making** could also have an impact on the procedure [19]. Furthermore, the **absence of a well-structured regulatory framework** or the **lack of a culture** that encourages compliance with procedural requirements [13] may be a burden in the research ethics review process. Additionally, the **lack of sufficient financial and human resources** [13] could stand as an obstacle to the scheduling of the RECs processes, causing possible delays in the assessment tasks. At the same time, the lack of human and financial resources is also an obstacle to the creation of committees with sufficient specialisation and diversity in expertise, causes difficulties in supporting the provision of training [13], and this has an impact on the structure of the committees. Another critical issue is that the excessive length of related documents, such as information sheets and necessary applications, can discourage participants, researchers, as well as REC members from engaging with them thoroughly, growing a checkbox-oriented mentality [10].

Bouhouita-Guermech et al. [17], in their examination of **challenges related to AI in research ethics**, identify **explainability** as a critical concern. Explainability issues are more evident in the context of Deep Learning (DL), where the use of extensive algorithms with data that have a broad range of interpretations raises difficulties in explaining the results of AI models [17, 20–22]. Consequently, the difficulty in explaining AI-generated

results raises **transparency** concerns, presenting challenges for research participants, as transparency is essential for ensuring that participants are fully informed and able to comprehend the various dimensions of a study [17, 23, 24]. The phenomenon of a lack of explainability and transparency is strongly linked to the so-called “black box” and leads to the **challenge of obtaining informed consent**, as participants may not have sufficient information to give their consent [17, 22, 25]. Additional key considerations for protecting participant rights and maintaining ethical research practices include **data protection** and **privacy** [17, 25]. It is crucial that researchers should take these issues into consideration when designing AI systems, as they are responsible for ensuring that privacy is maintained [17, 26]. The need to protect the privacy of participants can sometimes lead to dilemmas related to the progress of science. Studies that rely on more accurate data often involve higher risks to privacy [17, 27]. Moreover, other challenges related to research ethics that have an impact on the research ethics review are the **lack of awareness and knowledge of research ethics**, which generates a gap between research ethics and AI research, and the **lack of knowledge and training in new technologies**. It is observed that researchers involved in medical research have more understanding and knowledge on research ethics issues than computer scientists [17, 26, 28]. All of these issues, which are central to research ethics, also represent critical **challenges** that should be carefully considered **within the research ethics review** process. It is imperative to address these concerns and request additional information from the researchers during the ethics review process to ensure that studies are ethically and responsibly designed, thereby safeguarding research participants and users of AI systems.

3.2 Domain Specific Challenges

The literature review revealed that many studies articulate challenges and concerns regarding the use of AI across various fields. A significant portion focuses on the **health-care** sector, while other studies address concerns within the **education** sector, the **ICT** sector, and the **judicial** system. In general terms, the issues that raise concerns regarding the use of AI technologies in different fields and domains, converge in concepts such as **accountability**, **explainability**, **transparency**, **privacy**, **data protection**, **security**, and others that are also referred to as principles and requirements for the ethical use of AI.

In particular, studying the use of AI technologies in the health sector highlights the need for a deeper understanding of machine learning algorithms and prediction generation methods. This need derives from challenges related to **explainability** and **interpretability** in AI technologies related to the health sector [29, 30]. Another issue that poses a challenge for the use of AI technologies in this sector is **gender bias** in healthcare applications. It is important, in order to avoid leading to incorrect predictions, to use AI technologies that include data representative of all gender and sex differences and algorithms that do not reproduce existing biases or create new ones [29, 31]. In the healthcare sector, two further challenges related to the use of AI technologies are **accountability** and **safety**. The decisions produced by AI-driven clinical tools may be harmful to the patient, and moral accountability must be differentiated from legal liability [32].

The integration of AI in the judiciary introduces several ethical challenges that must be addressed to ensure fair and just outcomes. These challenges relate to **bias and fairness, transparency and accountability, privacy and data protection** [33]. Given that AI systems are fundamentally dependent on the data used for their training, the presence of hidden biases within data can substantially affect the performance and outputs of these models. Transparency and accountability pose another critical issue, particularly when the AI system operates with an opaque decision-making process, making it difficult to trace and understand how outcomes are generated. In the judiciary field, privacy and data protection present also significant concerns, as AI systems rely on large volumes of sensitive personal information, raising the risk of data breaches, unauthorised access, and misuse that could harm individuals [33].

In addition, similar challenges can be identified from the use of AI applications within the ICT sector. A typical example can be found in modern digital services and products which increasingly incorporate AI technologies, such as Large Language Models (LLMs). Their use raises issues related to **explainability, data privacy, safety, security, bias, and fairness** [34]. The use of AI technologies in education also requires particular consideration, since, as with health and justice, it is a very sensitive sector. The issues that arise in this sector in relation to the use of AI systems can be focused on the **accessibility and reliability** of AI systems [35]. The challenges arising from the use of AI technology can be partially addressed through the training and professional development of teachers.

4 Limitations and Future Research

This study attempts to identify, at a first level, challenges raised by the application of AI in different fields, and challenges related to research ethics review in general as well as regarding the application of AI technologies. This exploration aims to examine how the effectiveness of research ethics review processes can be enhanced for research involving AI technologies.

The examination of the issue has led to the creation of a framework for classifying the challenges, examining articles from Scopus, and drawing some primary conclusions. However, a deeper and more extensive review of the literature is needed, alongside empirical validation, that may lead to further enhancement of the classification framework. More in-depth research may result in the addition of other parameters that have not yet been taken into account. With regard to the domains in which AI is applied, it is important to examine emerging challenges across other fields and consider whether they converge with those already detected. Future research could also focus on and explore the potential challenges faced by RECs at different stages of the project lifecycle, in relation to the respective development phases of AI systems.

In light of the above, conducting empirical research with REC members is also valuable for reinforcing the existing classification framework and obtaining informed feedback. The members of the Research Ethics Committees that are involved and called upon to undertake the review of AI research proposals are the key stakeholders to assess the proposed classification framework and verify the findings of the literature. Structured interviews with REC members could bring to the forefront their perspective and their considerations, and may detect new challenges. The insights of the RECs offer the potential to lead to targeted policy recommendations.

We believe that this classification offers a fertile ground for better evaluation of AI research proposals. This perspective is grounded in the assumption that effectively addressing challenges requires first identifying the specific type of issue before developing targeted solutions. In this direction, a future study may focus, after reinforcing the classification framework through a more extensive literature research and empirical validation, on the formulation of policy proposals taking into consideration cultural and regulatory framework differences. Targeted policy recommendations are essential to contribute both to the research ethics review process and to the role of the Research Ethics Committees in a more efficient evaluation of research proposals related to AI technologies.

5 Conclusions

This study aimed to explore the emerging challenges posed by the widespread adoption of AI, focusing particularly on how these challenges influence research ethics review processes intended to uphold ethical and responsible research practices. These challenges broadly converge on core issues such as explainability, accountability, safety, privacy, data protection, and other critical dimensions that constitute the foundational principles for the ethical design, development, and deployment of AI systems. These principles are also fundamental requirements for achieving Trustworthy AI, ensuring that AI technologies are developed in a manner that is transparent, fair, and aligned with societal values. All of these challenges should be carefully considered during the evaluation of projects involving AI technologies to ensure ethical and responsible innovation. Simultaneously, it is essential to address challenges related to the structural and procedural aspects of RECs to ensure comprehensive and responsible oversight, thereby avoiding research that, although it seems ethically sound on the surface, may lack genuine ethical rigor. The creation of a framework for classifying the challenges related to AI technologies and their impact on research ethics and the research ethics review is a first approach to examine the field and to set a basis for addressing them through the formulation of policy recommendations. Further and deeper research in the literature and empirical validation of the classification framework through structured interviews with REC members is considered essential both for its strengthening and for its potential to lead to positive results, such as targeted policies.

Acknowledgement. This research is part of a PhD research and funded by the project CHANGER, Grant Agreement No 101131683, under European Commission Horizon WIDERA.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Floridi, L., et al.: AI4people – an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Mind. Mach.* **28**(4), 689–707 (2018). <https://doi.org/10.1007/s11023-018-9482-5>
2. European Commission: Directorate-General for Communications Networks, Content and Technology, Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji. Ethics guidelines for trustworthy AI. Publications Office (2019)
3. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems (2019)
4. Jobin, A., Ienca, M., and Vayena, E.: The global landscape of AI ethics guidelines (2019). <https://doi.org/10.1038/s42256-019-0088-2>
5. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. *Mind. Mach.* **30**, 99–120 (2020). <https://doi.org/10.1007/s11023-020-09517-8>
6. Hagendorff, T.: A virtue-based framework to support putting AI ethics into practice. *Philos. Technol.* **35**, 55 (2022). <https://doi.org/10.1007/s13347-022-00553-z>
7. Ziouvelou, X., Giouvanopoulou, K., Karkaletsis, V.: An integrated ethics framework for embedding values in AI. In: *The Leading Role of Smart Ethics in the Digital World*, pp. 277–289. Universidad de La Rioja (2024)
8. Laine, J., Minkinen, M., Mäntymäki, M.: Ethics-based AI auditing: a systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Inform. Manage.* **61**, 103969 (2024). <https://doi.org/10.1016/j.im.2024.103969>
9. Ghallab, M.: Responsible AI: requirements and challenges. *AI Perspect.* **1**(1), 1–7 (2019). <https://doi.org/10.1186/s42467-019-0003-z>
10. Dove, E.: Research ethics review. In: Laurie, G., et al. (eds.) *The Cambridge Handbook of Health Research Regulation*, pp. 177–186. Chapter, Cambridge University Press (2021)
11. Page, S.A., Nyeboer, J.: Improving the process of research ethics review. *Res. Integrity Peer Rev.* **2**, 14 (2017). <https://doi.org/10.1186/s41073-017-0038-7>
12. Ferretti, A., et al.: Ethics review of big data research: what should stay and what should be reformed? *BMC Med. Ethics* (2021). <https://doi.org/10.1186/s12910-021-00616-4>
13. Coleman, C.H., Bouësseau, M.C.: How do we know that research ethics committees are really working? The neglected role of outcomes assessment in research ethics review. *BMC Med. Ethics* **9**, 6 (2008). <https://doi.org/10.1186/1472-6939-9-6>
14. Kitchenham, B., Charters, S.M.: Guidelines for Performing Systematic Literature Reviews in Software Engineering. EBSE (2007)
15. Page, M.J., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* (2021). <https://doi.org/10.1136/bmj.n71>
16. Patton, M.Q.: *Qualitative Evaluation and Research Methods*, 2nd edn. Sage Publications, Newbury Park, CA (1990)
17. Bouhouita-Guermech, S., Gogognon, P., Bélisle-Pipon, J.C.: Specific challenges posed by artificial intelligence in research ethics. *Front. Artif. Intell.* **6**, 1149082 (2023). <https://doi.org/10.3389/frai.2023.1149082>
18. Laurie, G., Harmon, S.: Through the thicket and across the divide: successfully navigating the regulatory landscape in life sciences research. In: Cloatre, E., Pickersgill, M. (eds.) *Knowledge, Technology and Law*, pp. 121–136. Chapter, Routledge, London (2014)

19. Friesen, P., Yusof, A.N.M., Sheehan, M.: Should the decisions of institutional review boards be consistent? *Ethics Hum. Res.* **41**(4), 2–14 (2019). <https://doi.org/10.1002/eahr.500022>
20. Chassang, G., Thomsen, M., Rumeau, P., Sedes, F., Delfin, A.: An interdisciplinary conceptual study of artificial intelligence (AI) for helping benefit-risk assessment practices. *AI Commun.* **34**, 121–146 (2021). <https://doi.org/10.3233/AIC-201523>
21. Ienca, M., Ignatiadis, K.: Artificial intelligence in clinical neuroscience: methodological and ethical challenges. *AJOB Neurosci.* **11**, 77–87 (2020). <https://doi.org/10.1080/21507740.2020.1740352>
22. Jacobson, N.C., Bentley, K.H., Walton, A., Wang, S.B., Fortgang, R.G., Millner, A.J., et al.: Ethical dilemmas posed by mobile health and machine learning in psychiatry research. *Bull. World Health Organ.* **98**, 270–276 (2020). <https://doi.org/10.2471/BLT.19.237107>
23. Sedenberg, E., Chuang, J., Mulligan, D.: Designing commercial therapeutic robots for privacy preserving systems and ethical research practices within the home. *Int. J. Soc. Robot.* **8**, 575–587 (2016). <https://doi.org/10.1007/s12369-016-0362-y>
24. Grote, T.: Randomised controlled trials in medical AI: ethical considerations. *J. Med. Ethics* **48**, 899–906 (2021). <https://doi.org/10.1136/medethics-2020-107166>
25. Andreotta, A.J., Kirkham, N., Rizzi, M.: AI, big data, and the future of consent. *AI Soc.* **34** (7), 1715–1728 (2022). <https://doi.org/10.1007/s00146-021-01262-5>
26. Ford, E., Scarlett Shepherd, K.J., Lamie, H.: Toward an ethical framework for the text mining of social media for health research: a systematic review. *Front. Digit. Health* **2**, 592237 (2020). <https://doi.org/10.3389/fdgh.2020.592237>
27. Gooding, P., Kariotis, T.: Ethics and Law in Research on Algorithmic and Data-Driven Technology in Mental Health Care: Scoping Review. *JMIR Ment. Health* **8**, e24668 (2021). <https://doi.org/10.2196/24668>
28. Nebeker, C., Torous, J., Bartlett Ellis, R.J.: Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Med.* **17**, 137 (2019). <https://doi.org/10.1186/s12916-019-1377-7>
29. Mohammad Amini, M., Jesus, M., Fanaei Sheikholeslami, D., Alves, P., Hassanzadeh Benam, A., Hariri, F.: Artificial intelligence ethics and challenges in healthcare applications: a comprehensive review in the context of the European GDPR mandate. *Mach. Learn. Knowl. Extract.* **5**(3), 1023–1035 (2023). <https://doi.org/10.3390/make5030053>
30. Farah, L., Murris, J.M., Borget, I., Guilloux, A., Martelli, N.M., Katsahian, S.I.: Assessment of performance, interpretability, and explainability in artificial intelligence-based health technologies: what healthcare stakeholders need to know. *Mayo Clinic Proc.: Digit. Health* **1**(2), 120–138 (2023). <https://doi.org/10.1016/j.mcpdig.2023.02.004>
31. Fosch-Villaronga, E., Drukarch, H., Khanna, P., Verhoef, T., Custers, B.: Accounting for diversity in AI for medicine. *Comput. Law Secur. Rev.* **47**, 105735 (2022). <https://doi.org/10.1016/j.clsr.2022.105735>
32. Kumar, P., Chauhan, S., Awasthi, L.K.: Artificial intelligence in healthcare: review, ethics, trust challenges & future research directions. *Eng. Appl. Artif. Intell.* **120**, 105894 (2023). <https://doi.org/10.1016/j.engappai.2023.105894>
33. John, A.M., Aiswarya, M.U., Panachakel, J.T.: Ethical challenges of using artificial intelligence in judiciary. In: 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE), pp. 723–728. IEEE (2023). <https://doi.org/10.48550/arXiv.2504.19284>
34. Ray, P.P.: ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations, and future scope. *Internet of Things and Cyber-Physical Systems* **3**, 121–154 (2023). <https://doi.org/10.1016/j.iotcps.2023.04.003>

35. Flores-Vivar, J.-M., García-Peñalvo, F.-J.: Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4). *Comunicar* **31**(74), 37–47 (2023). <https://doi.org/10.3916/C74-2023-03>



Interpretability and the Measurement of Ethical Foundations in Artificial Intelligence

Miguel Houghton Torralba¹, Ziwei Shu²(✉) , Ramón-Alberto Carrasco² ,
and María Francisca Blasco López¹

¹ Department of Marketing, Faculty of Commerce and Tourism, Complutense University of Madrid, 28003 Madrid, Spain

² Department of Marketing, Faculty of Statistics, Complutense University of Madrid, Avenida Puerta de Hierro,1, 28040 Madrid, Spain
ziweishu@ucm.es

Abstract. With the rapid development of Artificial Intelligence (AI), its integration into decision-making processes across various sectors is accelerating. The demand for interpretability and ethical accountability has become more urgent than ever. This work explores the critical intersection of these two domains. It begins by examining the concept of interpretability in AI, then turns to the ethical foundations of AI. This work also examines how these intertwined concepts of interpretability and ethics are pivotal in advancing corporate social responsibility (CSR) by fostering transparency, enabling responsible governance, and addressing societal impacts such as algorithmic bias, job displacement, and environmental concerns. Integrating interpretability and ethics is essential for building transparent, accountable, and demonstrably ethically sound AI systems that proactively support robust CSR objectives and ensure profound alignment with human values and fundamental rights. This crucial integration helps create equitable opportunities for all, paving the way for a genuinely responsible and sustainable technological future that benefits society broadly and promotes inclusive growth.

Keywords: Artificial Intelligence · Corporate Social Responsibility · Ethics · Interpretability · Measurement

1 Introduction

Artificial intelligence (AI) has become increasingly integrated into decision-making processes across sectors, and the need for interpretability and ethical accountability has grown more urgent. However, the need for AI to be interpretable and explainable has become a central theme in both public and academic discourse regarding the ethical implications of this rapidly evolving technology. A major source of ethical concern in AI stems from the potential for biased algorithms, particularly because they often operate as a “black box”, making it difficult to understand how decisions are made. This opacity can lead to outcomes that reinforce or exacerbate existing social inequalities. Artificial neural networks, commonly used in AI to process data and learn patterns,

exemplify this issue. While they are powerful and capable of high accuracy in complex tasks, their internal decision-making processes are often difficult to interpret, raising important ethical and accountability concerns [1]. This opacity raises significant ethical questions, particularly when AI systems are used in contexts that demand transparency and accountability. As John-Mathews [2] notes, *“We consider two fundamental and related issues currently faced by AI development: the lack of ethics and interpretability of AI decisions.”* His observation underscores the intertwined nature of these challenges and highlights the need for more transparent, explainable AI systems to ensure ethical and responsible deployment.

This work aims to investigate the critical role of interpretability in understanding and, crucially, measuring the adherence of AI systems to their ethical foundations. The primary objective is to explore how various forms of interpretability can serve as mechanisms not only for explaining AI behavior but also for assessing its alignment with core ethical principles such as fairness, accountability, transparency, and non-maleficence. Furthermore, this paper aims to examine existing ethical frameworks and discuss the challenges and potential pathways for developing quantifiable metrics or qualitative assessment methods that leverage interpretability to evaluate the ethical posture of AI. Ultimately, the goal is to contribute to the discourse on building more responsible and ethically sound AI.

This work is structured as follows: Sect. 2 presents definitions of interpretability in AI and explores the ethical foundations of AI. Section 3 critically explores the role of AI interpretability and ethics in advancing corporate social responsibility (CSR). Finally, Sect. 4 summarizes the key findings of this work and outlines potential directions for future research.

2 Theoretical Framework

2.1 Interpretability in AI

Interpretability in AI refers to the ability to understand and explain how AI models make decisions. Interpretability ensures that AI models are transparent and understandable to humans, allowing stakeholders to scrutinize how decisions are made. This is especially critical in healthcare, finance, or law, where the consequences of decisions can significantly impact society and create controversy.

For instance, an AI diagnostic tool might flag a skin lesion as potentially malignant. Interpretability methods such as LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) can identify the specific visual features (e.g., irregular borders, color variegation) that most influenced the AI’s prediction. This allows a dermatologist to assess the reasoning behind the AI’s suggestion with their clinical expertise, thereby enabling ethical validation of the system’s output through human expertise before proceeding with clinical action such as a biopsy. Similarly, in finance, if a credit application is denied, explainability mechanisms are essential for regulatory compliance and procedural fairness. An interpretable model might provide reasons like “high debt-to-income ratio based on reported income X and debts Y” or “insufficient credit history length of Z months.” This information gives applicants actionable

feedback, ensuring transparency and minimizing perceptions of arbitrariness or bias—thereby quantifying ethical principles like fairness and accountability. In law, AI tools used for parole risk assessment might provide a score alongside contributing factors like “number of prior offenses” or “history of substance abuse.” Such disclosures not only enable defense attorneys to evaluate and challenge the model’s reasoning but also serve as a mechanism to qualify ethical concerns related to justice and due process.

The terms transparency and explainability are often used interchangeably with interpretability, though subtle differences exist:

- Transparency relates to the openness of the system’s structure, algorithms, and decision-making logic.
- Explainability focuses on providing reasons for an AI’s output in a human-understandable form.

About the transparency and explainability of AI systems—and considering the ethical requirements embedded in technical features—Balasubramaniam et al. [3] reveal that transparency is emphasized by nearly all organizations, with explainability regarded as a fundamental component of transparency. Specifically, explainability is closely linked to both transparency and the overall trustworthiness of AI systems. These findings underscore the importance of systematically defining explainability requirements as a critical step toward the development of transparent and trustworthy AI technologies. In practical deployments, methods to achieve interpretability range from using inherently simpler models (e.g., decision trees or linear regression) whose logic is directly inspectable, to employing post-hoc explanation techniques for more complex “black box” models. These techniques aim to enhance the interpretability of AI models by highlighting feature importance or simulating model behavior in specific cases. For instance, they can make results more understandable by linking a patient’s symptoms to the AI’s diagnostic suggestion. In this way, interpretability serves as a bridge between technical outputs and ethical evaluation, facilitating crucial assessments of whether the AI’s behavior is justifiable, equitable, and free from harmful biases.

To sum up, as AI systems develop rapidly, embedding interpretability at the core of their design is essential. Doing so not only facilitates regulatory compliance and risk mitigation but also serves a deeper ethical purpose: enabling AI to function as a tool for empowerment rather than a source of opaque control.

2.2 Ethical Foundations in AI

The ethical foundations in AI refer to the underlying moral principles and values that guide the development, deployment, and governance of artificial intelligence systems to ensure they promote human well-being, uphold individual rights, and avoid harm. These foundations include “intangible” characteristics such as transparency, fairness, accuracy, accountability, non-discrimination, and justice, which serve to align AI technologies with societal norms and ethical expectations. The concept of measurement of underlying ethics involves assessing how these ethical foundations are integrated into AI systems. For instance, measuring fairness in an AI-powered hiring tool might involve quantifying disparities in recommendation rates across different demographic groups using metrics like demographic parity or equalized odds, and then iteratively refining the model or

data to mitigate identified biases. Similarly, accountability in an autonomous driving system can be assessed through the robustness of its event data recorders and the clarity of pre-defined responsibility frameworks for decisions made in critical traffic scenarios.

It is essential to develop frameworks and metrics that can help organizations determine whether their AI systems align with ethical standards and societal values. By embedding these ethical foundations into AI design and policy frameworks, developers and policymakers aim to build systems that are not only technically effective but also socially responsible and trustworthy. For instance, transparency presents a key challenge in the practical implementation of AI ethics. While ethical principles such as fairness, accountability, and non-discrimination are increasingly emphasized in AI development, achieving true transparency often clashes with the complexity of modern machine learning models, particularly those based on deep learning. These models operate as “black boxes,” making it difficult for developers, regulators, and end-users to understand how specific decisions are made (i.e., low interpretability).

One proposed solution to this issue is the use of AI systems specifically designed to explain their decisions, therefore improving the interpretability of complex models by providing human-understandable rationales for their outputs, thus enabling developers, regulators, and end-users to gain insights into the decision-making process and fostering greater trust in AI systems. As Vainio-Pekka et al. [4] argue, explainable AI plays a crucial role in bridging the gap between ethical principles and technical implementation, enabling stakeholders to better understand, trust, and validate AI-driven outcomes. Vijayaraghavan and Badea [5] provide a valuable contribution by examining AI interpretability within the specific context of Machine Decision Making. Their research outlines various levels of interpretability, ranging from minimal transparency to fully explainable systems, and links these levels to different approaches for constructing Artificial Moral Agents. This framework helps clarify how the degree of interpretability directly affects the ethical capacity of AI systems, particularly in scenarios where moral reasoning and accountability are required, such as in AI judges adjudicating minor offenses or AI systems managing resource allocation in smart cities.

Furthermore, regulatory frameworks such as the European Union’s AI Act have proposed a risk-based approach that mandates higher standards of transparency, human oversight, and data governance for high-risk AI applications [6]. Similarly, international organizations like United Nations Educational, Scientific and Cultural Organization (UNESCO) emphasize the importance of inclusive, human-centered AI development, advocating for policies that ensure universal digital rights and cross-cultural equity [7]. These efforts reinforce the idea that integrating ethics into AI systems is not only a matter of technological design but also of broader societal responsibility.

To sum up, ethical foundations in AI are crucial not only for guiding the responsible design of intelligent systems but also for maintaining public trust. As AI continues to evolve, these principles and frameworks serve as essential guardrails, shaping the trajectory of AI in a way that aligns with human values and ensures fairness, accountability, and transparency in its deployment.

3 Role of AI Interpretability and Ethics in Advancing CSR

The integration of AI into the fabric of modern business operations presents opportunities and ethical challenges. As corporations increasingly leverage AI for everything from customer service and product development to strategic decision-making and operational efficiency, the traditional paradigms of CSR are being reshaped and expanded. The concept of CSR, a critical element of modern business practices, ensures that companies operate ethically, contribute to economic development, and improve the quality of life for their workforce, families, local communities, and society.

CSR and AI constitute a binomial that encompasses a broad range of activities, including environmental sustainability, ethical labor practices, philanthropy, and community engagement [8]. CSR and ethics are not merely optional; they contribute to a company's long-term success. Studies suggest that businesses with strong CSR initiatives enjoy higher customer loyalty, improved brand reputation, and better financial performance [9]. Ethical business conduct, on the other hand, involves adherence to moral principles, transparency, and compliance with legal standards [10].

An essential aspect of modern CSR is the role of AI interpretability and ethics in fostering transparency and building stakeholder trust. Modern CSR frameworks emphasize open communication and accountability to diverse stakeholders, including customers, employees, investors, regulators, and the public. By employing techniques that enable AI systems to explain their decision-making processes in human-understandable terms, companies demonstrate a commitment to transparency, aligning with the CSR principle of operating openly and honestly. As Floridi et al. [11] argue in their ethical framework for a good AI society, explicability, which encompasses both interpretability and transparency, is essential for ensuring AI is beneficial and ethically sound. When stakeholders understand why an AI system made a particular decision, their trust in the technology and the organization deploying it increases. This trust is invaluable, contributing to brand loyalty, employee engagement, and a positive public image—core objectives of CSR.

Furthermore, ethical decision-making in business, guided by principles such as honesty, integrity, and accountability, reinforces trust among stakeholders, including customers, employees, and investors [12]. Companies that uphold these standards not only avoid reputational and legal risks associated with unethical practices like corruption or exploitation but also strengthen the ethical foundation on which their AI systems are built. Therefore, the underlying ethics of AI should be seriously considered. Evolving in these concepts of CSR with AI, it is logical to admit that its interpretability, measurements, and ethics should be bonded. AI is transforming industries, but its development and ways to measurement raises significant ethical concerns and CSR obligations. Firms utilizing AI must balance innovation with ethical considerations, ensuring fairness, transparency, and accountability [13]. CSR in AI involves responsible AI development, mitigating biases, and addressing societal impacts. One major ethical issue in AI is bias in algorithms. AI systems trained on biased data can perpetuate discrimination in hiring, lending, and law enforcement [14]. Companies must adopt fairness measures and diverse datasets to reduce algorithmic biases. Transparency is another critical aspect; AI decision-making should be correctly interpreted and explained to build credence with users and regulators [15].

The scope of CSR in the AI era extends to specific societal and environmental impacts. One critical aspect of CSR is environmental responsibility. Companies adopt sustainable practices such as reducing carbon footprints, minimizing waste, and utilizing renewable energy sources to mitigate their environmental impact [16]. These efforts contribute to long-term ecological balance and demonstrate corporate accountability. Probably, the use of AI could help firms to improve the aforementioned reduction, but the point is whether this “advice” could be listened to by these organizations. Another crucial dimension of CSR is social responsibility, which includes fair labor practices, diversity, equity, and inclusion, and employee well-being. Ethical businesses ensure fair wages, safe working conditions, and equal opportunities for all employees [17]. A significant CSR concern related to AI is job displacement, as the widespread adoption of AI threatens millions of existing jobs. This necessitates corporate investment in workforce reskilling and education initiatives [18] as part of their social responsibility. Ethical businesses also engage in philanthropic activities, supporting community development, and these efforts can be amplified or complicated by AI’s societal integration.

Additionally, AI’s environmental impact, particularly in energy-intensive machine learning models, calls for sustainable AI development practices that should also be assessed carefully [19]. Fortunately, firms that prioritize ethical AI should foster trust and long-term success. Implementing responsible AI guidelines, promoting inclusivity, and ensuring accountability align with CSR principles and creating a positive societal impact. In an era of rapid AI advancement, businesses must proactively address accurate measurement and interpretation of ethical challenges to uphold their social responsibilities.

A significant challenge in this evolving landscape is ensuring the quality of CSR measurement in the context of AI. In the current environment where customers’ opinions are of fundamental importance, one of the major challenges companies face, regardless of sector, is ensuring the quality of CSR measurement in the context of AI. This quality can be evaluated by examining the gap between perceived value and expected value, as proposed by Chumpitaz Caceres and Paparoidamis [20]. The ability to accurately assess this gap depends heavily on the reliability of the data collected, which must adhere to established service quality standards. The interpretability of this data about CSR performance is as crucial as the interpretability of the AI systems themselves. Several studies offer methodological solutions to address these challenges. For instance, the SERVPERF scale has been shown to improve the measurement of perceived service quality by focusing solely on performance, reducing biases associated with expectations. Additionally, the use of the 2-tuple linguistic model developed by Herrera and Martínez [21] enables more precise handling of qualitative data, especially when dealing with subjective opinions expressed in natural language. This approach is particularly useful when integrating heterogeneous information from multiple sources or criteria, a need that is further addressed by the methodology proposed by Herrera et al. [22]. Furthermore, the work of Carrasco et al. [23] illustrates how these techniques can be effectively implemented in models designed to evaluate ethical dimensions in AI, offering a more nuanced and accurate assessment of CSR performance.

4 Conclusions

This work has explored the concepts of interpretability and ethical foundations in AI, as well as the essential connection between AI interpretability, ethics, and the advancement of CSR. It can be concluded that interpretability is not merely a required feature of AI systems but a fundamental enabler for operationalizing ethical principles and embedding them within robust and accountable CSR frameworks. By demystifying the “black box” of complex AI, interpretability fosters transparency, which is crucial for building and maintaining stakeholder trust—a cornerstone of effective CSR. It allows for the scrutiny necessary to identify and mitigate biases, ensuring that AI systems promote fairness rather than perpetuate existing social inequalities, directly addressing a major ethical concern in AI development.

Furthermore, the investigation highlighted how interpretability underpins ethical accountability. When AI systems produce unintended or harmful outcomes, the ability to understand their decision-making pathways is paramount for assigning responsibility, implementing corrective measures, and providing redress. This is vital for companies aiming to fulfill their CSR obligations, particularly as AI influences diverse societal domains, from employment and finance to environmental sustainability.

Companies must proactively address critical issues such as algorithmic bias, job displacement, and environmental impact to align AI development with broader societal values. By prioritizing interpretability and ethical accountability in their AI systems, organizations can build trust with stakeholders, minimize ethical and operational risks, and make meaningful contributions to social well-being. Ethics in AI is no longer just a regulatory requirement—it has become a cornerstone of long-term sustainability and organizational success. Companies that integrate ethical considerations into their AI strategies are more likely to foster innovation, enhance customer loyalty, and maintain a competitive edge in a value-driven market. In the broader context, ethically guided AI contributes to the creation of inclusive, fair, and resilient societies capable of thriving in the face of rapid technological change.

Acknowledgements. The authors would like to acknowledge the financial support provided by the Facultad de Estudios Estadísticos at the Universidad Complutense de Madrid.

Disclosure of Interests. The authors declare no competing interests relevant to the content of this article.

References

1. Pasquale, F.: *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press (2015)
2. John-Mathews, J.-M.: Some critical and ethical perspectives on the empirical turn of AI interpretability. *Technol. Forecast. Soc. Chang.* **174**, 121209 (2022)
3. Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkänen, K., Kujala, S.: Transparency and explainability of AI systems: from ethical guidelines to requirements. *Inf. Softw. Technol.* **159**, 107197 (2023)

4. Vainio-Pekka, H., et al.: The role of explainable ai in the research field of AI ethics. *ACM Trans. Interact. Intell. Syst.* **13**(4), 26:1–26:39 (2023)
5. Vijayaraghavan, A., Badea, C.: Minimum levels of interpretability for artificial moral agents. *AI and Ethics* **5**(3), 2071–2087 (2024)
6. European Commission. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>. Last accessed 25 Jan 2025
7. United Nations Educational, Scientific and Cultural Organization: Recommendation on the Ethics of Artificial Intelligence. <https://unesdoc.unesco.org/ark:/48223/pf0000380455>, Last accessed 26 Jan 2025
8. Carroll, A.B.: The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. *Bus. Horiz.* **34**(4), 39–48 (1991)
9. Kotler, P., Lee, N.: *Corporate Social Responsibility: Doing the Most Good for Company and Your Cause*. Wiley (2005)
10. Crane, A., Matten, D., Glozer, S., Spence, L.: *Business Ethics: Managing Corporate Citizenship and Sustainability in the Age of Globalization*, 5th edn. Oxford University Press (2019)
11. Floridi, L., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Mind. Mach.* **28**(4), 689–707 (2018)
12. Trevino, L.K., Nelson, K.A.: *Managing Business Ethics: Straight Talk about How to Do It Right*. John Wiley & Sons (2016)
13. Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. In: Floridi, L. (ed.) *Ethics, Governance, and Policies in Artificial Intelligence*. PSS, vol. 144, pp. 5–17. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-81907-1_2
14. O’Neil, C.: *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (2016)
15. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning (2017). [arXiv:1702.08608](https://arxiv.org/abs/1702.08608)
16. Kramer, M.R., Porter, M.: Creating shared value. *Harvard Business Review* (2011)
17. Parmar, B.L., Freeman, R.E., Harrison, J.S., Wicks, A.C., Purnell, L., De Colle, S.: Stakeholder theory: the state of the art. *Acad. Manag. Ann.* **4**(1), 403–445 (2010)
18. Brynjolfsson, E.: *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (2014)
19. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for modern deep learning research. *Proc. AAAI Conf. Artif. Intell.* **34**(09), 13693–13696 (2020)
20. Chumpitaz Caceres, R., Paparoidamis, N.G.: Service quality, relationship satisfaction, trust, commitment and business-to-business loyalty. *Eur. J. Mark.* **41**(7/8), 836–867 (2007)
21. Herrera, F., Martínez, L.: A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Trans. Fuzzy Syst.* **8**(6), 746–752 (2000)
22. Herrera, F., Martínez, L., Sánchez, P.J.: Managing non-homogeneous information in group decision making. *Europ. J. Operat. Res.* **166**(1), 115–132 (2005). <https://doi.org/10.1016/j.ejor.2003.11.031>
23. Alberto Carrasco, R., Blasco, M.F., García-Madariaga, J., Herrera-Viedma, E.: A fuzzy linguistic RFM model applied to campaign management. *Int. J. Interact. Multimed. Artif. Intell.* **5**(4), 21 (2019). <https://doi.org/10.9781/ijimai.2018.03.003>



The Mediating Effect of Job Crafting in the Relationship Between Organizational Commitment and Organizational Citizenship Behavior and Its Ethical Implications

Melissa Eugenia Treviño Serna^(✉) , Rosalba Martínez Hernández ,
and Juan Carlos Yáñez Luna 

Universidad Autónoma de San Luis Potosí, 78000 San Luis Potosí, Mexico
melissa.trevino@uaslp.mx

Abstract. The present study analyzes the mediating role of job crafting in the relationship between organizational commitment and organizational citizenship behavior in hotels. The data analysis was made with a sample of 137 hotel employees from 10 different hotels, PLS-SEM was used to run the different tests to corroborate our hypothesis. Findings proved the mediating effect of job crafting between organizational commitment and organizational citizenship behavior, as well as the relationship between organizational commitment, organizational citizenship behavior and job crafting.

This study contributes to understanding how these variables interact with each other and their effects in the hospitality industry, as well as providing evidence to comprehend ethical behavior in technology-mediated environments, aligning with international frameworks on responsible business conduct.

Keywords: Organizational Commitment · Job Crafting · Organizational Citizenship Behavior · Hospitality Industry

1 Introduction

The study of job crafting and its relationships in different contexts has been increasing in the last years due to the continuous changes occurring globally, such as the increasing tendency to the employees being more proactive in their duties [1]; the increasing interest in the use of creativity by employees as a means for crafting their jobs and fulfill their duties [2]; the importance of facilitating job crafting in service business in order to contribute business success [3]; the need of job designers in updating their methodologies to include job crafting practices in the hospitality business [4] and the focusing of companies on developing task competences in their employees and providing them with the required resources to aid them in working well and independently [5].

Companies are taking great interest in learning about the type of individuals who are working under their lead, in terms of their performance, organizational change adaptation [6], motivation [7], commitment [4, 8] and other job-related aspects [9, 10] and

mechanisms which they use to cope with the demands of their jobs as well as developing policies related to ethical business practices.

Not to mention, each company has the duty of safeguarding their employees about the uses of technology in their daily processes, as is stated by Anwar & Deliana [11], the incorporation of new technologies in the hospitality industry has improved the quality of their services and increased the efficiency of their business processes, enhancing their engagement. Technologies such as Internet of Things (IoT), wearables, sensors and actuators, have revolutionized the way in which hotels offer their services by providing customized experiences, namely: smart hotel rooms with controlled items, location-aware services to improve customer interaction, satisfaction and contributes to a better understanding and responding of the customer needs [12], the use of cloud computing, virtual and augmented reality enable the stakeholders participation and can lead to co-create value processes as well as providing immersive and interactive customer experiences, and the implementation of AI by making personalized recommendations with the use of customer data, can lead to drive business success, optimize tasks and improve decision-making [13].

As mentioned by Wynn & Lam, there is a variation regarding the use of digital technologies in the hospitality sector, whilst some companies incorporate them within their strategical planning and core processes, others lack of an integrated strategy and policy, which is reflected in reporting inconsistencies, disproportionated maintenance, as well as security issues [14].

In increasingly digital workplaces, job crafting becomes not only an engagement strategy, but also a space for ethical decision-making, regarding the responsible use of technology which enables identifying and recruiting skills sets and competencies, monitoring changes in organizational culture and individual attitudes, along with the implications of convincing the workforce to leave their comfort zone.

According to Gallup Consulting in his report *State of the Global Workplace* when managers are engaged, the employees are more likely to be engaged and this can lead to positive outcomes in organizations such as: profits, retention rates and better customer service [15].

Concurring with this, DNV & United Nations Global Compact state in his report: *IMPACT: Transforming Business, Changing the World* that companies are investing in more resources to improve performance as well as working to enclose responsibility for performance into relevant corporate functions, as a consequence of the rising understanding of the connection between business performance, sustainability and its impact into core business strategies [16].

The International Labor Organization emphasizes responsible business conduct through the MNE declaration, an instrument that provides direct guidance to enterprises (multinational and national) on social policy, inclusive, responsible and sustainable workplace practices [17].

Supporting this initiative, the OECD has developed the *Guidelines for Multinational Enterprises on Responsible Business Conduct*, addressing various aspects by covering different business-related aspects, including self-regulatory practices that might help companies in developing and implementing adequate internal controls and ethical programs preventing any act of corruption and bribery [18].

However, not only the creation of these types of regulations and guidelines is enough.

Considering this, Abuzaid & Ghadi [19] state the importance of prioritizing ethical conduct within the organization by promoting and enhancing proactive behaviors and aligning the organizational and individual values. As well as the *Código de Integridad y Ética Empresarial* [20] published by the Consejo Coordinador Empresarial addresses the importance of the fight against corruption within companies and promotes principles of integrity and business ethic which companies can incorporate within their internal policies.

The present study addresses the relationship between organizational commitment and organizational citizenship behavior through job crafting in the hospitality industry and will analyze the direct effect of organizational commitment on job crafting, the effect of job crafting on the organizational citizenship behavior in hospitality industry; the effect of organizational commitment on organizational citizenship behavior and the mediating effect of job crafting in the relationship between organizational commitment and organizational citizenship behavior.

This research not only provides empirical support for these relationships but also introduces elements for an ethical understanding of work environments in which technology is highly prevalent and is essential for consolidating a framework for studying social-technological systems and their ongoing relationship with organizational practices.

The findings of this work contribute significantly to the ethical design of sociotechnical systems in organizations.

2 Literature Review

2.1 Organizational Commitment

One of the first antecedents about Organizational Commitment, was presented in 1960 with the side-bet theory developed by Becker [21], which states that by making side-bets and linking external interests with a specific activity, commitment would be generated as a result. Years later, Meyer and Allen [22] in 1984 tested Becker's theory through a study in which they concluded that this theory mainly focused on measuring the affective commitment and suggested the use of other indicators that could measure the individuals perceptions about organizational commitment without being based on their bets.

Later that year they presented a study in which they introduced the distinction between affective and continuance commitment. As well as they suggested the normative commitment as a third component. In which they proposed that the affective commitment denotes an emotional attachment, identification and involvement of the individual with the organization; the continuance commitment bases on the associated costs that individuals incurred of leaving the organization, and the normative commitment reflects an individual perceived obligation of remaining in the organization. Based on this, identifying the type and degree of commitment in the organizations, it would be possible to know what kind of employees they have, their motivations to remain in the organization, as well as to predict the outcomes based on their level of commitment.

2.2 Job Crafting

Karasek in 1979 [23], developed an organizational stress model stating the importance of redesigning organizational processes and granting workers a certain degree of freedom in making decisions that contributes in the reduction of mental stress.

However, the concept of job crafting and their practices were introduced at the end of the 80's, considering internal and external changes and their impact on the organization. Kulik et al. [24] examined the ways in which the job characteristics and personal attributes of the employees converge with time, taking into account the job characteristics theory. With the passage of time, new studies surged showing the changes that employees made in their jobs by their own volition and that contributed to the organization, without a formal planning [25].

In 2001, Wrzesniewski & Dutton defined job crafting as the physical and cognitive changes that individuals made in their tasks or the job relational limits by their own volition which are related to individual task performance, social characteristics and individual thinking [26, 27]. As well as focuses specifically on job redesign by the employees by modifying job-related aspects with the aim to improve and adapt them accordingly to their needs, skills and preferences [28].

Based on the previous definitions, we must infer that job crafting can be defined as the workers' initiative to modify aspects related to their daily chores with the purpose of customizing and make them more bearable, this entails a job redesign by the individual which enables them to make decisions without losing sight of the essence and importance of fulfilling their duties in order to achieve the organizational objectives.

2.3 Organizational Citizenship Behavior (OCB)

The Organizational Citizenship Behavior (OCB) is a construct which has been present since the late 1930's, during the "human relationship era" [29], and through time, many authors who have investigated it, have been introducing diverse definitions related to the OCB, which have evolved through the passage of time.

Katz and Khan in 1966 defined the OCB as actions which are not subjected to a specific position, contributed to the organization functioning, without impacting task performance [30]. Years later Organ in 1988 defined OCB as an individual behavior, which was not stated in the job description, however, its contribution relied on the promotion of well-functioning of the organization, and was considered as a personal choice to make [31]. In the following years several authors defined OCB as behaviors which go beyond or above job formal requirements [32], as a synonym of extra-role performance [33, 34]. As useful behaviors which go beyond the expected job tasks [35]. As extra-role behaviors [32–38]. Many of these definitions have concurred that the focal point of the OCB is benefitting the organization, and also mention behaviors non established in the job formal descriptions.

With the previous conceptualizations about OCB, we can define it as non-formal established actions in the job descriptions, which individuals perform to benefit the organization by their own volition and with a proactive vision.

2.4 Organizational Commitment and Job Crafting

The positive relationship between organizational commitment and job crafting has been analyzed in the works of Qi et.al [39] in which they found that affective commitment has a positive effect on job crafting in employees of a manufacturing company; in their study Cheng et al [40] concluded that job crafting has a positive influence on the organizational commitment of tour leaders of travel agencies, and Iqbal [41] concurred with this positive influence on bank employees. Based on these antecedents regarding the relationship between organizational commitment and job crafting, the following hypothesis is presented (Fig. 1):

Hypothesis 1: Organizational Commitment has a direct effect on job crafting.



Fig. 1. Hypothesis 1

2.5 Job Crafting and Organizational Citizenship Behavior (OCB)

To know about the relationship between job crafting and Organizational Citizenship Behavior (OCB), the works of Mahmoud [42], Peral & Geldenhuys [43], and Maruhum Hutapea [44] et al. were analyzed and their findings confirm that job crafting has a positive influence on OCB [42, 44]; task and cognitive crafting are related positively with individual OCB [43]. Based on these antecedents regarding the relationship between job crafting and OCB, the following hypothesis is presented (Fig. 2):

Hypothesis 2: Job crafting has a direct effect on OCB.



Fig. 2. Hypothesis 2

2.6 Organizational Commitment and Organizational Citizenship Behavior (OCB)

The works of Ibrahim & Aslinda; Hossain and Aulia Soleha [45–47]s how evidence of the relationship between organizational commitment and organizational citizenship

behavior, in which their findings proved that organizational commitment is positively related to OCB [45]; there is a significant and strong relationship between organizational commitment and OCB [46]; and organizational commitment has a significant positive effect on OCB [47]. Based on these antecedents, the following hypothesis is presented (Fig. 3):

Hypothesis 3: Organizational commitment has a direct effect on OCB.



Fig. 3. Hypothesis 3

2.7 The Mediating Effect of Job Crafting

The works of Li, Ingusci et al. and Peral & Geldenhuys [43, 48, 49] shown evidence regarding the mediating effect of job crafting on different variables, in which their findings proved that the mediating effect of job crafting in leader- member exchange and affective commitment [48]; the partial mediating effect of job crafting in self-efficacy and work performance and the total mediating effect in self-efficacy and OCB [49]; and the mediating effect between extraversion and individual OCB [43].

Based on these antecedents, the following hypothesis is presented (Fig. 4):

Hypothesis 4: Job Crafting has a mediating effect on the relationship between Organizational Commitment and Organizational Citizenship Behavior.



Fig. 4. Hypothesis 4

3 Methods

The present study is cross-sectional, with a quantitative, descriptive and correlational approach. The population consisted of employees of hotels in the municipality of Ciudad Valles, in the State of San Luis Potosí, México, with a total sample of 137 hotel employees from 10 hotels.

4 Measures

The data collection was conducted through a questionnaire, in which 10 items were used to measure organizational commitment by using the instrument developed by Meyer et al. [50] with a 7 point Likert Scale (1 = completely in disagree -7 totally agree).

For the measurement of job crafting were used 4 items of the instrument developed by Tims et al. [51] with a 7 point Likert scale (1 = never-7 Always).

As for the OCB, its measurement was made with 4 items of the instrument developed by Spector et. al [52] with a 7 point Liker Scale (1 = never-7 Always).

The data analysis was performed on the software SMART- PLS 4.

The methodological framework is based on exploring complex relationships underlying the variables previously mentioned and their interactions in the hospitality industry. Therefore, the methodology chosen for developing this study is based on the PLS-SEM [53] due to it centers on the analysis of total variance and focuses on measuring the model and its structure.

5 Results

5.1 Validity and Reliability Test

According to the validity and reliability test, the results in the construct of job crafting show .848 in Cronbach's Alpha, .855 in composite reliability and .590 in AVE, indicating a good level of internal consistency in the construct's indicators, as well as a good level of composite reliability and that the construct is validated through its indicators.

The results regarding the construct of OCB, show a good level of internal consistency with Cronbach's Alpha of .805, a good level of composite reliability of .814, and an AVE of .517, indicating the validation of the construct through its indicators.

As for the construct of organizational commitment, the results show an excellent level of internal consistency with a Cronbach's alpha of .912, an adequate level of composite reliability with .926 and an AVE of .512, indicating the validation of the construct through its indicators. All these results are shown in Table 1.

Table 1. Construct reliability and validity

Construct	Cronbach's alpha	Composite reliability (rho_a)	AVE
Job Crafting	0.848	0.855	0.590
Organizational Citizenship Behavior	0.805	0.814	0.517
Organizational Commitment	0.912	0.926	0.512

5.2 Discriminant Validity

In terms of the discriminant validity of the constructs, the coefficients are OCB to job crafting .716, organizational commitment to job crafting .568 and Organizational commitment to OCB .656, indicating that there is not conceptual overlapping among each of the constructs. As is shown in Table 2.

Table 2. Discriminant Validity

	Heterotrait-monotrait ratio (HTMT)
Organizational Citizenship Behavior < ->Job Crafting	.716
Organizational Commitment < ->Job Crafting	.568
Organizational Commitment < ->Organizational Citizenship Behavior	.656

5.3 Collinearity

Table 3 provides the results of the collinearity test, in which it can be appreciated that the VIF levels are below 3.0, indicating an acceptable collinearity level, as well as suggesting that the constructs are independent of each other.

Table 3. Inner Model Values

	VIF
JC ->OCB	1.487
OC->JC	1.000
OC->OCB	1.487

5.4 Direct, Indirect and Total Effects

Table 4 presents the path coefficients, including direct and indirect effects. The results show that the path coefficient of JC ->OCB (.502) denotes a strong positive relationship between these constructs; as well as in the path OC->JC (.572); and in the path OC->OCB the result was a substantial effect. As for the indirect effects, the path OC->JC ->OCB (.287) shows that job crafting has a moderate mediator effect in the relationship between organizational commitment and OCB.

As for the total effects, the results of Table 5 show that job crafting has a total effect on OCB (.502), organizational commitment strongly influences job crafting (.502), and OCB (.660).

Table 4. Direct and Indirect Effects

	Path coefficients
JC ->OCB	.502
OC->JC	.572
OC->OCB	.372
OC->JC ->OCB	.287

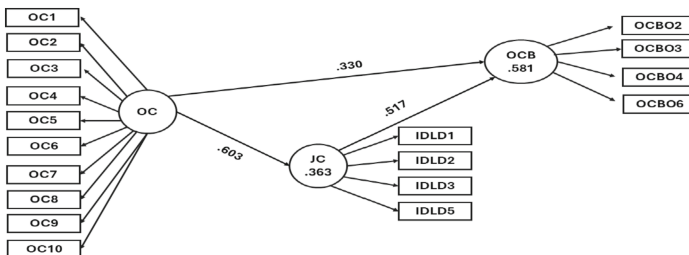
Table 5. Total Effects

	Path coefficients
JC ->OCB	.502
OC->JC	.572
OC->OCB	.660

On Table 6, the results of the f-square values show the contribution of each of the constructs, in which it can be appreciated that OCB has a substantial effect on job crafting and organizational commitment, as well as job crafting has a strong effect on organizational commitment. With these results we observe the contributions of each of the constructs (Fig. 5).

Table 6. f-square values

	f-square
JC->OCB	.429
OC->JC	.487
OC->OCB	.236

**Fig. 5.** Model research

6 Conclusions

Based on the results previously stated, we conclude that organizational commitment has a positive and direct effect on job crafting and OCB, with this Hypothesis 1 and 3 are supported; job crafting has a strong positive influence on OCB, therefore Hypothesis 2 is sustained, and regarding the mediating effect on job crafting in the relationship between organizational commitment and OCB, the Hypothesis 4 is proven.

The findings of this study may serve as a solid foundation for hotels, companies across various sectors, decision-makers, and future researchers by providing a reference framework for understanding the impact of organizational commitment on proactive behaviors within organizations. This empirical evidence enables these stakeholders to assess and reflect on the ethical implications of analyzing the benefits of job crafting and organizational citizenship behaviors (OCB), not only in terms of employee commitment but also in achieving broader organizational objectives. Furthermore, the importance of fostering proactive and ethical work environments is emphasized, as such environments facilitate organizational adaptation in an increasingly demanding and dynamic context.

In this regard, organizations must ensure that proactive initiatives do not translate into excessive workloads or unreasonable expectations for employees, always respecting their well-being and autonomy. Job crafting and OCB should be encouraged within a framework of fairness, organizational justice, and respect for human dignity. Taking these ethical dimensions into account is essential to prevent well-intentioned practices from resulting in employee exploitation or undue pressure disguised as commitment or productivity.

The cross-sectional design of the current study limits the ability to infer causal relationships over time. Future longitudinal research is recommended to explore how organizational commitment influences job crafting and OCB in the long term. Practically, organizations can implement ethical job crafting actions by offering employees autonomy to redesign tasks, setting clear boundaries to avoid overload, and promoting open communication. For example, allowing staff to adjust roles based on strengths, while monitoring burnout risks, fosters engagement ethically. Future studies could compare industries or regions to assess how cultural or sector-specific factors shape these dynamics, offering broader insights for customized interventions.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

- Schachler, V., Eppele, S.D., Clauss, E., Hoppe, A., Slemp, G.R., Ziegler, M.: Measuring job crafting across cultures: Lessons learned from comparing a German and an Australian sample. *Front. Psychol.* **10**, 1–13 (2019). <https://doi.org/10.3389/fpsyg.2019.00991>
- Tian, W., Wang, H., Rispen, S.: How and when job crafting relates to employee creativity: the important roles of work engagement and perceived work group status diversity. *Int. J. Environ. Res. Public Health* **18**, 1–17 (2021). <https://doi.org/10.3390/ijerph18010291>
- Bala, S.R., Bhattacharya, S., Thakur, M.: Job crafting, performance, and the mediating role of gratitude: a study among the sales service personnel of tourist industry. *India. ASEAN J. Hosp. Tour.* **20**, 53–66 (2022). <https://doi.org/10.5614/ajht.2022.20.2.04>

4. Abbas, T.M., Mansour, N.M., Elshawarbi, N.N.M.A.: Job crafting and organizational commitment: the mediating role of person-job fit in the food and beverage sector. *Tour. Hosp. Manag.* **29**(3), 319–333 (2023). <https://doi.org/10.20867/thm.29.3.1>
5. Nwanzu, C.L., Babalola, S.S.: Employee job crafting behaviour and its antecedents: a study of psychological safety, psychological autonomy, and task competence. *Empl. Responsib. Rights J.* (2024). <https://doi.org/10.1007/s10672-024-09520-6>
6. Petrou, P., Demerouti, E., Xanthopoulou, D.: Regular versus cutback-related change: the role of employee job crafting in organizational change contexts of different nature. *Int. J. Stress Manage.* **24**(1), 62–85 (2017). <https://doi.org/10.1037/str0000033>
7. Kim, S.H., Kim, M., Holland, S.: Effects of intrinsic motivation on organizational citizenship behaviors of hospitality employees: the mediating roles of reciprocity and organizational commitment. *J. Hum. Resour. Hosp. Tour.* **19**, 168–195 (2020). <https://doi.org/10.1080/15332845.2020.1702866>
8. Dash, S.S., Vohra, N.: The leadership of the school principal: Impact on teachers' job crafting, alienation and commitment. *Manag. Res. Rev.* **42**, 352–369 (2019). <https://doi.org/10.1108/MRR-11-2017-0384>
9. Kataria, A., Garg, P., Rastogi, R.: Do high-performance HR practices augment OCBs? The role of psychological climate and work engagement. *Int. J. Product. Perform. Manag.* **68**, 1057–1077 (2019). <https://doi.org/10.1108/IJPPM-02-2018-0057>
10. Luu, T.T.: Linking authentic leadership to salespeople's service performance: the roles of job crafting and human resource flexibility. *Ind. Mark. Manag.* **84**, 89–104 (2020). <https://doi.org/10.1016/j.indmarman.2019.06.002>
11. Anwar, F.A., Deliana, D., Suyamto, S.: Digital transformation in the hospitality industry: improving efficiency and guest experience. *Int. J. Manage. Sci. Inform. Technol.* **4**(2), 428–437 (2024). <https://doi.org/10.35870/ijmsit.v4i2.3201>
12. Mercan, S., Akkaya, K., Cain, L., Thomas, J.: Security, privacy and ethical concerns of IoT implementations in hospitality domain. In: *IEEE 13th International Conference on Internet of Things (iThings 2020)* (2020)
13. Dang, T.D., Nguyen, M.T.: Systematic review and research agenda for the tourism and hospitality sector: co - creation of customer value in the digital age. *Futur. Bus. J.* **9**, 1–15 (2023). <https://doi.org/10.1186/s43093-023-00274-5>
14. Wynn, M., Lam, C.: Digitalisation and IT strategy in the hospitality industry. *Systems* **11**(10), 501 (2023). <https://doi.org/10.3390/systems11100501>
15. Gallup Consulting: State of the Global Workplace. Washington, DC (2024)
16. DNV & United Nations Global Compact: IMPACT Transforming Business, Changing the World. Høvik, Norway (2024)
17. International Labour Organization: Tripartite Declaration of Principles concerning Multinational Enterprises and Social Policy (MNE Declaration). <https://www.ilo.org/ilo-department-sustainable-enterprises-productivity-and-just-transition/areas-work/tripartite-declaration-principles-concerning-multinational-enterprises-and>
18. OECD: OECD Guidelines for Multinational Enterprises on Responsible Business Conduct (2023). <https://doi.org/10.1787/81f92357-en>
19. Abuzaid, A.N., Ghadi, M.Y.: The effect of ethical leadership on innovative work behaviors: a mediating–moderating model of psychological empowerment, job crafting, proactive personality, and person–organization fit. *Administrat. Sci.* **14**(9), 191 (2024). <https://doi.org/10.3390/admsci14090191>
20. Consejo Coordinador Empresarial (CCE): Código de Integridad y Ética Empresarial. México (2017)
21. Becker, H.S.: Notes on the concept of self-representation. *J. Am. Psychoanal. Assoc.* **6**, 32–40 (1960). <https://doi.org/10.1177/000306515700500207>

22. Meyer, J.P., Allen, N.J.: Testing the “side-bet theory” of organizational commitment: some methodological considerations. *J. Appl. Psychol.* **69**, 372–378 (1984). <https://doi.org/10.1037/0021-9010.69.3.372>
23. Karasek, R.A.: Job demands, job decision latitude, and mental strain: implication for job redesign. *Adm. Sci. Q.* **24**, 285–308 (1979)
24. Kulik, C.T., Oldham, G.R., Hackman, J.R.: Work design as an approach to person-environment fit. *J. Vocat. Behav.* **31**, 278–296 (1987). [https://doi.org/10.1016/0001-8791\(87\)90044-3](https://doi.org/10.1016/0001-8791(87)90044-3)
25. Benner, P., Tanner, C., Chesla, C.: *Expertise in Nursing Practice: Caring. Springer Publishing, Clinical Judgement and Ethics* (1996)
26. Kantén, P.: The antecedents of job crafting: perceived organizational support, job characteristics and self-efficacy. *Eur. J. Bus. Soc. Sci.* **3**, 113–128 (2014)
27. Tims, M., Derks, D., Bakker, A.B.: Job crafting and its relationships with person-job fit and meaningfulness: a three-wave study. *J. Vocat. Behav.* **92**, 44–53 (2016). <https://doi.org/10.1016/j.jvb.2015.11.007>
28. Tims, M., Bakker, A.B., Derks, D., van Rhenen, W.: Job crafting at the team and individual level: implications for work engagement and performance. *Gr. Organ. Manag.* **38**, 427–454 (2013). <https://doi.org/10.1177/1059601113492421>
29. Ocampo, L., et al.: A historical review of the development of organizational citizenship behavior (OCB) and its implications for the twenty-first century. *Personnel Rev.* **47**(4), 821–862 (2018). <https://doi.org/10.1108/PR-04-2017-0136>
30. Bateman, T.S., Organ, D.W.: Job satisfaction and the good soldier: the relationship between affect and employee “citizenship.” *Acad. Manag. J.* **26**, 587–595 (1983). <https://doi.org/10.2307/255908>
31. Podsakoff, P.M., MacKenzie, S.B., Paine, J.B., Bachrach, D.G.: Organizational citizenship behaviors: a critical review. *J. Manage.* **26**, 513–563 (2000)
32. Ackfeldt, A.L., Coote, L.V.: An Investigation Into The Antecedents of Organisational Citizenship Behaviours. *Young* 217–222 (2000)
33. Chen, Z.X., Francesco, A.M.: The relationship between the three components of commitment and employee performance in China. *J. Vocat. Behav.* **62**, 490–510 (2003). [https://doi.org/10.1016/S0001-8791\(02\)00064-7](https://doi.org/10.1016/S0001-8791(02)00064-7)
34. Blakely, G.L., Srivastava, A., Moorman, R.H.: The effects of nationality work role centrality, and work locus of control on role definitions of OCB. *J. Leadersh. Organ. Stud.* **12**, 103–117 (2005). <https://doi.org/10.1177/107179190501200109>
35. Spector, P.E., Fox, S.: Theorizing about the deviant citizen: an attributional explanation of the interplay of organizational citizenship and counterproductive work behavior. *Hum. Resour. Manag. Rev.* **20**, 132–143 (2010). <https://doi.org/10.1016/j.hrmr.2009.06.002>
36. Ball, G.A., Trevino, L.K., Sims, H.P., Jr.: Just and unjust punishment: influences on subordinate performance and citizenship. *Acad. Manag. J.* **37**, 299–322 (1994). <https://doi.org/10.2307/256831>
37. Yeh, C.M.: The relationship between tourism involvement, organizational commitment and organizational citizenship behaviors in the hotel industry. *Tour. Hosp. Manag.* **25**(1), 75–93 (2019). <https://doi.org/10.20867/thm.25.1.10>
38. Ma, E., Wang, Y.-C., Xu, S.T., Wang, D.: Clarifying the multi-order multidimensional structure of organizational citizenship behavior: a cross-cultural validation. *J. Hospit. Tourism Manage.* **50**, 83–92 (2022). <https://doi.org/10.1016/j.jhtm.2021.12.008>
39. Qi, J., Li, J., Zhang, Q.: How organizational embeddedness and affective commitment influence job crafting. *Soc. Behav. Pers.* **42**, 1629–1638 (2014). <https://doi.org/10.2224/sbp.2014.42.10.1629>
40. Cheng, J.-C., Chen, C.-Y., Teng, H.-Y., Yen, C.-H.: Tour leaders’ job crafting and job outcomes: the moderating role of perceived organizational support. *Tour. Manag. Perspect.* **20**, 19–29 (2016). <https://doi.org/10.1016/j.tmp.2016.06.001>

41. Iqbal, Q.: Job-crafting and organizational commitment: person-job fit as moderator in banking sector of Pakistan. *Int. J. Manag. Account. Econ.* **3**, 837–851 (2016)
42. Mahmoud, H.: Job crafting and work involvement as a mediator to promote head nurses ' organizational citizenship behavior at Mansoura University Hospitals. *Int. J. Nurs. Didact.* **7**, 1–9 (2017)
43. Peral, S.L., Geldenhuys, M.: The indirect relationship between personality and performance through job crafting behaviour. *SA J. Ind. Psychol.* **46**, 1–12 (2020). <https://doi.org/10.4102/sajip.v46i0.1715>
44. Maruhum Hutapea, T., Indradewa, R., Yanuar, T., Syah, R., Fajarwati, D.: HR practices, employment engagement, job crafting, organizational citizenship behavior affect task performance (Bekasi city banking company). *J. Multidiscip. Acad.* **5**, 234–239 (2021). <https://doi.org/10.51971/joma.v5n3.058106021>
45. Ibrahim, M.A., Aslinda, A.: Relationship between organizational commitment and organisational citizenship behavior (OCB) at government-owned corporation companies. *J. Public Adm. Gov.* **3**, 35 (2013). <https://doi.org/10.5296/jpag.v3i3.4379>
46. Hossain, M.M.: Organizational citizenship behavior and organizational commitment among clinical nurses in Bangladesh. *Open J. Nurs.* **10**, 693–704 (2020). <https://doi.org/10.4236/ojn.2020.107049>
47. Aulia Soleha, D.: The role of organizational commitment in mediating the influence of self-efficacy and work-life balance on organizational citizenship behavior. *Int. J. Econ. Bus. Issues.* **03**, 23–33 (2022)
48. Li, J.: The mediating roles of job crafting and thriving in the LMX-employee outcomes relationship. *Japanese J. Adm. Sci.* **28**, 39–51 (2015). <https://doi.org/10.5651/jaas.28.39>
49. Ingusci, E., et al.: Self-efficacy and work performance: the role of job crafting in middle-age workers. *Int. J. Bus. Soc.* **20**, 533–551 (2019)
50. Meyer, J.P., Allen, N.J., Smith, C.A.: Commitment to organizations and occupations: extension and test of a three-component conceptualization. *J. Appl. Psychol.* **78**, 538–551 (1993). <https://doi.org/10.1037/0021-9010.78.4.538>
51. Sora, B., Caballer, A., García-Buades, E.: Validation of a short form of job crafting scale in a spanish sample. *Span. J. Psychol.* (2018). <https://doi.org/10.1017/sjp.2018.52>
52. Spector, P.E., Bauer, J.A., Fox, S.: Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior : do we know what we think we know ? *J. Appl. Psychol.* **95**, 781–790 (2010). <https://doi.org/10.1037/a0019477>
53. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Black, W.C., Anderson, R.E.: *Multivariate Data Analysis*. Cengage Learning. EMEA, Hampshire, U.K. (2018). <https://doi.org/10.1002/9781119409137.ch4>



On the Current (Im) possibility of Achieving Public Value Through the EU's Digital Strategy: An Ethics Method to Seek a “Collectual” Equilibrium

Stefano Calzati^(✉) 

Joint Research Centre, Brussels, Belgium
stefano.calzati@ec.europa.eu

Abstract. This paper has two aims, one theoretical and one practical. On a theoretical level, the goal is to highlight the criticalities and the ultimate impossibility to achieve public value by/through digital technologies, based on the current regulatory framework of the European Union's (EU) digital strategy. On a practical level, the goal is to redress such criticalities by advancing a complementary practice to the EU's normative framework, which takes the form of a “problem-opening” transdisciplinary method for questioning, exposing, and redressing the effects of digital technologies' use in complex scenarios. The method, discussed and exemplified in the paper, was tested from 2021 to 2024 at the Delft University of Technology in a course titled “Ethics for the Data-driven City”.

Keywords: EU · Digital Strategy · Public Value · Data Ethics · Education

1 The EU's “People-Centric” Digital Transformation

When it comes to governing the digital transformation, the EU's digital strategy aims to pursue a “people-centric” approach. This idea remains quite fuzzy in its characterization across documents, but in principle, the approach aims to safeguard human and data subjects' rights while balancing economic competitiveness with social inclusiveness, digital sovereignty, and environmental sustainability [1–5]. As such, the EU's “people-centric” approach emphasizes both the individual and collective dimensions of the digital transformation – or what I call a “collectual” (collective + individual) equilibrium – insofar as “people-centric” is concerned as much with the rights of the individual as with principles pertaining to society as a whole.

Notably, the Declaration of Digital Rights and Principles (DDRP), which represents the compass of the EU's digital strategy, defends “a European way for the digital transition, putting people at the center.” More specifically, what the DDRP does is to pin down six principles: (1) preserve people's rights; (2) support solidarity and inclusion; (3) ensure freedom of choice; (4) foster democratic participation; (5) increase safety, security, and empowerment of individuals; and (6) promote sustainability. For the present discussion, it is noteworthy that these principles are equally split between a half (1, 3,

5) focusing on the individual and the other half (2, 4, 6) pertaining to society. In this regard, the DDRP does seek a balance between the individual and collective dimensions affected by the digital transformation, focusing as much on the fundamental rights of the single person as on population-level principles irreducible to the former nor to their sum. Yet, the extent to which the EU's digital strategy manages to realize, de facto, such a "collectual" equilibrium has been subject to criticism.

Today, the Charter of Fundamental Rights (CFR) constitutes the polar star of the EU's "people-centric" approach. The rationale is to tackle the adversary effects of digital technologies by protecting citizens from the potential infringement of their fundamental rights [6]. While this is certainly relevant, recent studies have shown the limits of a right-based only standpoint toward the digital transformation. This is so because fundamental rights mainly focus on the individual, ultimately overlooking the population-level effects of the digital transformation [7] as well as its potential societal harm [8]. For instance, [7] notes that such an individualistic focus does not account for the relational nature of data and the consequent trade-off effects that data re-use involving two subjects might have on unaware third parties. Similarly, [8] suggests taking inspiration from environmental law for tackling potential societal harms caused by the digital transformation – such as, the erosion of the functioning of the rule of law – which can be neither accounted for nor mitigated by the current individualistic focus. Hence, to the extent to which a right-based approach does constitute the fundamental baseline to citizens' autonomy before the digital transformation, it might not be sufficient to protect Europeans as a whole [9].

To rebalance this scenario, it is necessary to complement the current EU's normative framework with the recognition of the systemic complexity of the digital transformation. This means moving away from prioritizing either certain actors (e.g., private or public), principles (oftentimes, economic competitiveness over social inclusion or environmental sustainability), or factors (mostly legal and technical specifics over soft skills and competences) to rather enact a balanced arena whereby actors, factors, and principles are tackled as co-dependent aspects of a whole sociotechnical ecosystem. This idea can be best disentangled in three respects: i) by elaborating on the concept of *systemic fairness* [10]; ii) by exploring a *non-normative* understanding of public value [11, 12]; iii) by questioning the *nature* and *effects* of technology beyond normative (do vs don't) and axiomatic (good vs bad) axes [13–15].

1.1 Sociotechnical Ecosystems Governed Through Systemic Fairness

An ecosystem is characterized by homeostasis, that is, the balanced interaction between biotic and non-biotic elements within a given environment. As an emergent behavior, this balance cannot be studied by isolating either elements or interactions; rather, it must be studied in its entirety. While largely related to the natural world, the notion of ecosystem has also been applied to other settings, such as the digital transformation [16, 17]. From here, a sociotechnical ecosystem comes to encompass the factors and actors contributing, through shared principles and agreed rules and mechanisms, to manage digital initiatives in a balanced way over time, i.e., fairly. In this context, fairness represents the meta-principle guiding the trading off among different (data) interests in view of the overall sustainability of the sociotechnical ecosystem. On this point, [10] notes that in documents part of the EU's digital strategy, fairness constitutes a structuring

principle “linked to principles such as proportionality and other procedural dimensions of a balancing exercise involving rights and interests.” In other words, moving beyond a reductionist conceptualization [18], fairness, as understood by the EU, entails and demands the enactment of a systemic equilibrium.

Already today, national, and supranational legal frameworks are in place for establishing rules and mechanisms for disentangling individual and collective interests concerning the access and (re)use of (personal) data (e.g., GDPR). This signals the extent to which alongside and beyond a dimension *de iure*, to achieve fairness *de facto* is a context-based task to be operationalized and assessed on a rolling basis. From a theoretical perspective, fairness shall reflect the diversity of interests of all actors involved in a given ecosystem [19]; yet, from a practical perspective, fairness shall also enable, especially in conflicting cases, the principled synthesis (not necessarily the sum) of all actors’ interests [20] in view of the whole ecosystem’s balance. Hence, beyond its normativity, the operationalization of fairness can never be established once and for all, but requires ongoing assessment, disentangling individual and collective dimensions.

1.2 Public Value as a Fractal

This understanding links to discussions on public value. We repeatedly hear that the digital transformation, especially *by/through* the public sector, is meant to benefit the collectivity at large. The disjunction *by/through* is crucial here in that it highlights and disentangles the nexus between public value(s) and the digital transformation. Public value theory has emerged as an alternative to the efficiency-based rationale of New Public Management [21]. In particular, public value theory in connection with the digital transformation puts the stress, not much on economics-inspired performances, but on the ability of the public sector to meet the expectations of citizens through more or less collaboratively designed services [22]. For instance, [23] identified six dimensions of public value that the digitalization of the public sector is expected to yield: improved public services; improved administrative efficiency; open government (OG) capabilities; improved ethical behavior and professionalism; improved trust and confidence in government; and improved social value and well-being. In a similar vein, research [24] on the release of data as open government data (OGD) showed how such a release yields the creation of public value (singular), from both governmental organizations and citizens, in the forms (plural) of transparency, accountability, and collaboration.

In this context, some authors [25] distinguish public value/s emphasizing either the *effects* that the public sector achieve *through* the digital transformation (*public values* in the plural) or the set of *principles* that guide the public sector *toward* the digital transformation (*public value* in the singular). However, the singular-plural pair can hardly be solved along the principle-effect axis, or by assuming that public value is the sum of shared public values. This is so because the pair is symptomatic of the tension between the individual and collective dimensions of the digital transformation.

In this regard, a stream of literature informing public value theory advances a more radical position, notably that any discussion on public value shall move beyond both its institutional confinement to the public sector as well as its sole legal focus [11]. Such a non-normative perspective aims to acknowledge the full roundedness and contradictory nature of human experience [12] leading to understand public value as a relational

concept emerging at the crossroad of subjective evaluation and society as a whole [11]. What public value stands for, then, is not an objective concept, but an individually and collectively constructed frame – or a *fractal*. In connection to that, work in behavioral ethics [26] highlighted the multifaceted nature of people’s decision-making whenever they are confronted with moral dilemmas. This means not only that most of individuals’ actions fall in a grey area that resists a clear-cut discernment in terms of “good vs. bad”, but also that individuals are *inconsistent agents* when it comes to understand what they value, insofar as the same subject might well endorse mutually exclusive values at once [27]. Hence, a regulatory normative focus on the individual and on public value, especially in relation to the digital transformation, are insufficient to realize a “collectual” equilibrium and must be accompanied with a real-world exploration of how people use digital technologies, why, and with which consequences.

1.3 Technology as a *Pharmakon*

Works in philosophy of technology can be very helpful in this regard, by advancing a non-normative understanding of technology and a non-axiomatic framing of its effects. Notably, [14] – and [13] before him – elucidated the pharmacological nature of technology, where *pharmakon* refers, in its Greek etymological meaning, to the double-sidedness of technology as both poison and antidote. Along the same line, [15] identified six laws of technology, the first of which reads: “Technology is neither good nor bad, *nor* is it neutral [emphasis added]”. This means, contrary to the refrain that either good or bad effects of technology depend on the use one makes of it, that, at all times, technology can have *both* positive and negative (un)intended consequences whenever it is applied in a given context. This is so because its uses generate value-laden entanglements that cannot be taken apart, neither with regard to the individual-collective nexus, nor concerning these two dimensions considered in isolation. As an example, the ethnographic work conducted by [27] exposes the conundrum of subjects willing to take environmentally friendly decisions, for instance concerning their commuting by electric vehicles, and how these decisions, once impacted by the pandemic, tangled with non-environmentally friendly behaviors in order to preserve the subjects’ personal health and that of their families (a good example of the fractal nature of public value).

The effects of the uses of technologies in a given context, then, is a multidimensional non-zero-sum game, whereby goodness or badness are, more often than not, superposed concepts from the perspective of the single subjects and the collective they are part of. This implies, in turn, that the public value/s pair is irreducible to objectivist categorizations and demands, instead, ongoing negotiation and hierarchization of its various (“collectual”) facets and downsides at once. There cannot be, for instance, openness (e.g., of data) and transparency, without acknowledging and tackling inevitable closures and opacities at individual and collective levels. Similarly, a digital service designed to promote social inclusion might achieve that for certain people and not for others; or it might be inclusive under certain conditions, but then result exclusive for the same people in later occurrences. Let’s see how these insights can be operationalized.

2 Transdisciplinarity for the Digital Transformation

From the theoretical discussion above, the following points can be summarized:

- the EU’s digital strategy emphasizes both the individual and collective dimensions impacted by the digital transformation – yet the digital strategy’s operationalization fails to strike a “collectual” equilibrium. Beyond a normative frame, it is necessary to approach the digital transformation as a sociotechnical ecosystem whose balance requires an ongoing exercise of systemic fairness.
- the disjuncture between public value and public values cannot be solved only normatively because at stake is both the striking of a “collectual” equilibrium (i.e., public value as a fractal) and the acknowledgment of the inconsistency of individual and collective agents in their decisions (i.e., the conflicting hierarchization of what they value).
- the nature and effects of technology cannot be reduced axiomatically in terms of good vs bad, nor be normatively ruled in terms of dos vs don’ts. As a pharmakon, technology produces in-context value-laden entanglements and un/intended consequences that require situated analysis.

To operationalize these three points, I will discuss what I call a “problem-opening” method, that is, the concretization of a transdisciplinary practice able to question and cut across epistemological boundaries, resist any privilege point of reference, and configure an ongoing multidimensional analysis of the effects of digital technologies in complex scenarios. In the remaining of the article, I will describe the application and results of such a method in an educational setting.

2.1 Teaching Data Ethics

Works Together with a colleague, from 2021 to 2024 we designed and taught a Master elective course in “Ethics for the Data-driven City” at the Delft University of Technology. Elsewhere [28], we already discussed the rationale of the course; here, it suffices to recuperate its main theoretical tenets and practical design.

The course (for a total of 14 classes over two months and a half) was based on three theoretical pillars: 1) data as “assemblages” part of a sociotechnical ecosystem [29]; 2) the city as a complex system showing emerging behaviors irreducible to computation [30, 31]; 3) ethics as an ongoing reflective practice synthesizing individual and collective interests and values [32, 33]. Based on this, we developed a transdisciplinary “problem-opening” method (as opposed to “problem-solving”), enabling students to a) identify the (“collectual”) ethical tensions emerging whenever digital technologies are implemented in an urban context; b) systemically reflect on these tensions; c) apply this understanding to the investigation of an independently chosen case study.

If we speak of transdisciplinarity it is because, as [34] notes, transdisciplinarity “concerns the dynamics engendered by the action of several levels of reality at once.” This is key to enable students to grasp the non-zero-sum complexity of the ethical dilemmas at stake whenever digital technologies are used in context. From here, we envisioned the exploration of the double-sidedness of tech-related principles – such as, trust, openness, transparency, privacy, inclusion – as well as of the unavoidable emergence of

un/intended consequences at individual and collective levels. For instance, [35] show that trust in/through digital technologies shall be best explored as an entangled concept – ‘dis/trust’ – which accounts for the duplicitous co-presence (and needed balance) of its own two opposites. The same goes for personal data: as [36] claims, at stake shall be the fundamental awareness that there is no clear-cut way to discern once and for all whether a certain set of data contains personal data or not, as these are two complementary features. Along a similar line, [37] contends that the opening up of governmental data has become an increasingly complex and contested affair. From a pedagogical point of view, it is precisely this complexity to become the focus of attention: they do not happen “by chance”, they are systemic and require transdisciplinary tackling.

To this aim, the course was organized as a *research-and-design workshop*. The first three classes were introductory ones, with each touching upon one of the three theoretical pillars of the course, enriched with real-life examples, including a visit to the TU Delft’s living lab The Green Village.

Concerning the workshop-style of the course, each class revolved around a given topic – e.g., health, mobility, carbon footprint, education, governance – in connection with the development, implementation, and use of digital technologies. Before each class students had to read two compulsory articles. In the first part of each class, a student presented and discussed one of the two articles, making sure that the presentation was linked to a current real-life example and/or to the second article. Then, through moderated debate, the ethical “collectual” tensions, the double-sidedness of principles, and the emergence of un/intended consequences related to the topic were brought to light. The second part of each class was dedicated to discussing students’ advancement toward the realization of their final assignment, either individually or collectively. Indeed, the final assignment, which relates more closely to the research-and-design part of the course, was meant to originate through an ongoing mentored process, rather than as an output to be submitted after the end of the course.

The research-and-design part was arranged as follows. Since the first class, students were asked to identify a possible case study linking a chosen digital technology – a platform, a service, a device – to a given urban setting. To facilitate their choice, we encouraged students to focus on TU Delft campus as their “city”, since developments, implementations, and testing of digital technologies are ongoing on site and students are more likely to be familiar with these projects. Once the case study was identified, the research part comprised of three steps: i) *framing* (e.g. how the case study is presented; how it works; who leads it/who is involved; its function, targets, and goals); ii) *data-value entanglements* (e.g. what are the principles mobilized and foregrounded, as well as those sidelined; which data are collected and which ones are overlooked; how the “collectual” equilibrium is negotiated); iii) *open-ended scenarios* (e.g. mapping expected and/or obtained results; possible un/intended – both positive and negative – consequences; reflection on how to redress these and with which further consequences). For each step, students were encouraged to reach out to the involved stakeholders (e.g., developers, implementers, or users).

Concerning the design part, the assignment consisted of two outputs: i) a project journal; ii) a (final) artefact. Since we were interested in the learning process as much as in the final result, we asked students to keep track from the end of class 1, by means of a

project journal, of their own reflections about the research part of their work, jotting down ideas and advancements towards the realization of their final artefacts. The journal had to be in writing, but we let students free to add other media formats (drawings, photos, images, etc.). By artefact, we meant the realization of a (simple) physical or digital object – e.g., a model, a chart, a video, an app mockup – that exposed and/or redressed the identified ethical tensions in the chosen case study. Last, we asked students to write a short “statement” (1000 words max) describing the artefact and addressing the following four points: a) what they focused on; b) why; c) which ethical tensions emerged; d) how these tensions were explored through the designed artefact.

2.2 Example

Works In this last section, I will provide one example of an artefact realized by a student with a background in geomatics, who enrolled in the course during the 2023–2024 academic year. As a case study, the student decided to focus on the digital platform of the main campus library, which allows students to book allocated study rooms over a certain desired timeslot, for either individual or group study sessions. In principle, the platform is meant to coordinate (and maximize) the use of these public spaces, based on the students’ individual needs. The student framed the case study as follows:

the platform monitors the occupancy through in-room sensors. When a student books through the platform, the desired timeslot becomes unavailable and cannot be booked by others. A screen placed next to the entrance of the rooms displays the current and upcoming bookings. During especially busy periods, such as exam sessions, it is not unusual for the rooms to be completely booked throughout the entire day, sometimes weeks in advance. This online platform aims to (supposedly) improve the “efficiency” of the physical space through measurable quantities. However, solutions that use assumingly “objective” data to portray and affect human behavior can often miss specificities, irregularities or deviant behavior.

To unpack the underlying principles of efficient use and (data) objectivity, the student carried out an ethnographic study (Fig. 1) during which he observed the occupancy and use of three study rooms over a period of twelve hours. Then, he compared his observations with the data on the platform. This is how he described the process:

On paper, I gathered notes parallel to each of the rooms’ booking timelines, indicating the actual time of arrival and departure of students as well as the size of the groups that would use the rooms. The observed behavior varied, from larger groups punctually following the booking times to single users staying in the rooms across multiple booked timeslots. The contrasting behavior of users was not reflected in the simple timeline shown on the platform’s website, in fact it often directly contradicted it. While the reasons for these contradictions could be many – such as an intentional overbooking by the same user, an indifference towards the bookings shown on the room’s display or a no-show booking slot being used by someone else – they clearly showed a disconnect between daily use of the study space and the platform’s data-based portrayal.

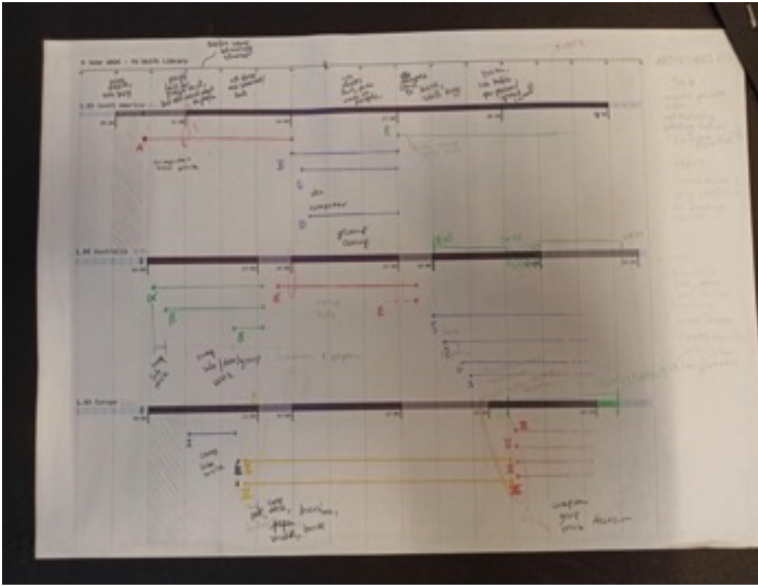


Fig. 1. Student's ethnographic notes (image copyrights ©Stefano Calzati).

On the one hand, these observations provide a good exemplification of the disjunction between norm and (practical) use. In fact, they highlight the tension between the smooth rationalizing principle underpinning the functioning of the platform and the, more or less, individual(istic) use made of the physical rooms. In a way, then, the questioning of the objectivity (and correctness) of the data appearing on the platform is at stake, alongside the questioning of whether such objectivity can amount to a fair outcome. On the other hand, such a norm-use disjunction brings efficiency and collective benefit on divergent paths, and this is so, as the student noted, *regardless* of the good or bad intentions on the users (an issue confined to virtue ethics). In other words, the fair use of the rooms, especially during busy exam periods, is at odd with the mere adherence to a rationalizing principle in terms of empty/occupied, especially when booking occurs weeks in advance. Not only the striking of a “collectual” equilibrium is a contextual ongoing exercise, but values at stake (e.g., the library as a public space, making available public spaces for temporary “private” use, as well as the possibility of ignoring, complying with, disrupting, or taking advantage of the booking system) are to an extent irreconcilable (e.g., one might occupy a room because it is empty, although it appears as booked, and this could hardly be objectively labeled as good or bad from either an individual or collective perspective). This signals that, alongside or even beyond a regulatory framework (i.e., the digital platform), effective “collectual” equilibrium requires practical self-organization and assessment (e.g., by letting a student to study in the room *even if* others are already present, depending on a fair negotiation).

To expose these tensions, the student worked on a qualitative representation of the platform data *vis à vis* his own observations, translating (i.e., materializing) his findings into the final artefact. These were his words:

To add the additional layer of user behavior, I wanted to employ a method of representation that differed from the two-dimensional digital timeline used by the platform and that would allow for more complexity. To this end, I explored the use of strings, braids and knots as a way of physically representing the gathered data. I briefly researched the *kipu*, the cord and knot-based writing system used by the Inkas, to investigate how information can be represented through this physical medium. The final artifact is not itself a *kipu* but is in part influenced by it.

Figures 2 and 3 show the final artefact. The main horizontal knotted cord represents the whole duration of the student's ethnography (12 h) and the (supposed) start and end of the booking periods for each of the observed rooms (knots at regular intervals). The three pending threads represent the three observed rooms over the considered time, while the vertical pendants stand for the stay of each student in the room (where each knot is a 15-min period). As the student wrote, “the artifact can be read from left to right, both visually and tactically, and allows the reader to look for the discrepancies between the positions of the knots and of the pendant cords, indicating the discrepancies between the bookings and the room use.”

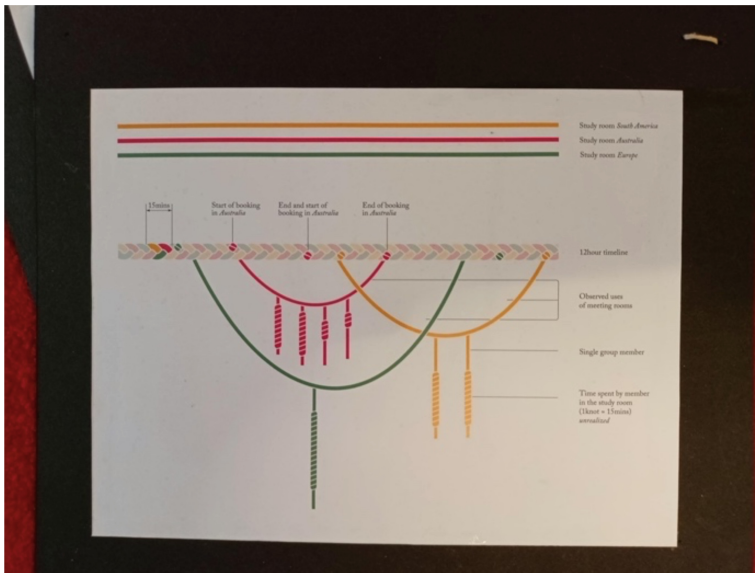


Fig. 2. Explanation of the artefact (image copyrights ©Stefano Calzati).

To this, the student added a further reflection on the discrepancy between his own artefact and the materialization of his collected data, intended as sociotechnical assemblages. In fact, while he was “satisfied” by his final result, the student also pointed out that “I was unable to find ways to include additional information that I recorded, such as the type of work that was conducted by the different groups.” In a way, then, not only the student signaled the exploratory nature of his work, demanded by the complexity of the scenario he was facing, but the artefact itself remains an open-ended materialization,

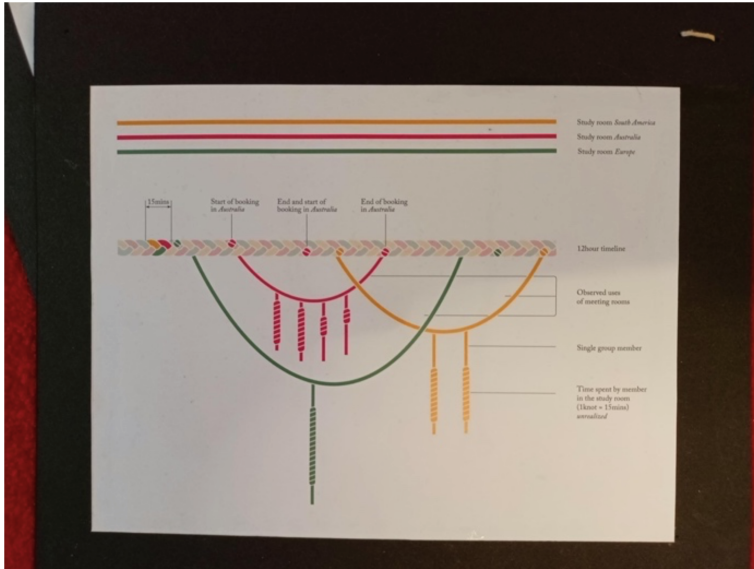


Fig. 3. The physical artefact (image copyrights ©Stefano Calzati).

not much in terms of interpretation, but concerning the impossibility of reciprocally fully mapping one set of data (digital records from the platform) and the student's own observations on to the material affordances proper to the artefact. Inconsistencies, one might say, pertain not only to the ethical decisions of individuals and groups, but also to the incommensurable nature of the physical vs. the digital.

3 Conclusion

The argument developed in this paper was double-sided. On a theoretical level, it aimed to show that, although the EU claims to pursue a “people-centric” digital strategy that strikes a balance between individual and collective rights and principles, what is being realized fails to strike such a balance, or a “collectual” equilibrium. This is because the normative framework set in place by the EU cannot account for a) the sociotechnical complexity of the digital transformation; b) the fractal (non-normative) understanding of public value, together with the inherent inconsistency of individual and collective ethical choices; c) the pharmacological (non-axiomatic and non-normative) nature of technology and its effects.

From here, the practical part of the article discussed the main tenets of a transdisciplinary practice aimed at complementing the EU's normative framework. In particular, such a practice was enacted in a course titled “Ethics for the Data-driven City”, designed and taught at TU Delft. Based on three pillars – data as sociotechnical assemblages; the city as a complex system; ethics as an ongoing reflective practice – the course developed a “problem-opening” method encouraging students to explore the inevitable value-laden entanglements and un/intended consequences of the use of digital technologies in a given

setting, in order to expose/redress such “collectual” ethical tensions through their materialization into an artefact. While the course was limited in time and scope, its rationale can be applied to different sectors as well as scaled up involving diverse stakeholders, contributing to a more comprehensive vision on the impact of technological “solutions” in delivering public value.

References

1. European Commission: Declaration of digital rights and principles (2022). <https://digital-strategy.ec.europa.eu/en/library/declaration-european-digital-rights-and-principles>.
2. European Parliament and Council: AI act (2024). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf
3. European Parliament and Council: Data act (2022). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN>
4. European Parliament and Council: Data governance act (2020b). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868>.
5. European Parliament and Council: A European strategy for data (2020a). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066>.
6. Aizenberg, E., Van Den Hoven, J.: Designing for human rights in AI. *Big Data Soc.* (2020). <https://doi.org/10.1177/2053951720949566>
7. Viljoen, S.: A relational theory of data governance. *Yale Law J.* **131**, 573 (2021)
8. Smuha, N.A.: Beyond the individual: governing AI’s societal harm. *Internet Policy Rev.* (2021)
9. Calzati, S., van Loenen, B.: Beyond federated data: a data commoning proposition for the EU’s citizen-centric digital strategy. *AI Soc.* **21**, 1–13 (2023)
10. Rochel, J.: Ethics in the GDPR: a blueprint for applied legal theory. *Int. Data Priv. Law* **11**(2), 209–223 (2021)
11. Meynhardt, T.: Public value inside: what is public value creation? *Int. J. Public Adm.* **32**(3–4), 192–219 (2009)
12. Talbot, C.: Paradoxes and prospects of ‘public value.’ *Public Money Manage.* **31**(1), 27–34 (2011)
13. Derrida, J.: *Dissemination*. Athlone Press, London (1981)
14. Lemmens, P.: An interview with Bernard Stiegler. *Krisis* **1**, 33–41 (2011)
15. Kranzberg, M.: Technology and history: Kranzberg’s laws. *Technol. Cult.* **27**(3), 544–560 (1986)
16. Jarke, M., Otto, B., Ram, S.: Data sovereignty and data space ecosystems. *Bus. Inf. Syst. Eng.* **61**(5), 549–550 (2019)
17. van Loenen, B., et al.: Towards value-creating and sustainable open data ecosystems: a comparative case study and a research agenda. *JeDEM* **13**(2), 1–27
18. Wong, P.-H.: Democratising algorithmic fairness. *Philos. Technol.* **33**(2), 225–244 (2020)
19. Healey, P.: *Collaborative planning*. McMillan, London (1997)
20. Innes, J.E., Booher, D.E.: A turning point for planning theory? Overcoming dividing discourses. *Plann. Theory* **14**(2), 195–213 (2015)
21. Panagiotopoulos, P., Klievink, B., Cordella, A.: Public value creation in digital government. *Gov. Inf. Quart.* **36**(4), 101421 (2019). <https://doi.org/10.1016/j.giq.2019.101421>
22. Jørgensen, T.B., Bozeman, B.: Public values. *Adm. Soc.* **39**(3), 354–381. <https://doi.org/10.1177/0095399707300703> (2007)
23. Twizeyimana, J.D., Andersson, A.: The public value of E-Government: a literature review. *Gov. Inf. Q.* **36**(2), 167–178 (2019)

24. Benmohamed, N., Shen, J., Vlahu-Gjorgievska, E.: Public value creation through the use of open government data in Australian public sector: a quantitative study from employees' perspective. *Gov. Inf. Quart.* <https://doi.org/10.1016/j.giq.2024.101930> (2024)
25. MacLean, D., Titah, R.: A systematic literature review of empirical research on the impacts of e-Government: a public value perspective. *Pub. Adm. Rev.* **82**(1), 23–38. <https://doi.org/10.1111/puar.13413> (2022)
26. Prentice, P.: Teaching behavioral ethics. *J. Leg. Stud. Educ.* **31**(2), 325–365 (2014)
27. Pink, S., et al.: Automation in electric vehicle futures. *Mobilities* **12**, 1–19 (2024)
28. Calzati, S., Ploeger, H.: Problem-solving? No, problem-opening! A method to reframe and teach data ethics as a transdisciplinary endeavour. *Big Data Soc.* (2024). <https://doi.org/10.1177/20539517241270687>
29. Kitchin, R., Lauriault, T.: Big data, new epistemologies and paradigm shifts. *Big Data Soc.* (2014). <https://doi.org/10.1177/2053951714528481>
30. Portugali, J.: Complexity, cognition and the city. Springer (2011)
31. Bettencourt, L.M.: Cities as complex systems. In: Furtado, B.A., Sakowski, P.A.M.E., Tóvolli, M.H.E. (eds.). *Modelling complex systems for public policies* (pp. 217–236). Ipea (2015)
32. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* **1**, 501–507 (2019)
33. Zigon, J.: Can machines be ethical? On the necessity of relational ethics and empathic attunement for data-centric technologies. *Soc. Res. Int. Quart.* **86**(4), 1001–1022 (2019)
34. Nicolescu, B.: Towards transdisciplinary education. *J. Transdisciplinary Res. Southern Afr.* **1**(1), 5–16 (2005)
35. Duenas-Cid, D., Calzati, S.: Dis/Trust and data-driven technologies. *Internet Policy Rev.* **12**(4), 1–23 (2023)
36. Purtova, N.: Health data for common good: Defining the boundaries and social dilemmas of data commons. In: Adams, S., Purtova, N., Leenes, R. (eds.) *Under Observation: The Interplay Between eHealth and Surveillance*, pp. 177–210. Springer, Berlin (2017)
37. Bates, J.: Open government data and the neoliberal state. <https://blogs.lse.ac.uk/impactofsocialsciences/2014/10/02/open-governmentdata-and-the-neoliberal-state/>



Inclusive Governance of Artificial Intelligence: Towards an Ethical Framework for Neurodivergence

Elisabet Bolarín-Miró¹  , María Mercedes Carmona Martínez¹ ,
and Asunción López-Arranz² 

¹ Catholic University of San Antonio of Murcia (UCAM), Murcia, Spain
ebolarin@alu.ucam.edu

² University of A Coruña, A Coruña, Spain

Abstract. Artificial intelligence (AI) is increasingly present in daily life and institutional decisions, yet its rapid development has outpaced ethical frameworks that address cognitive diversity. Neurodivergent individuals—including those with autism, ADHD, dyslexia, and Tourette syndrome—are systematically excluded by AI systems designed around neuronormative assumptions. This article examines how current AI technologies, from automated hiring to customer service and educational platforms, reproduce and amplify structural discrimination by failing to accommodate diverse cognitive profiles. A critical review of the European legal and policy framework as of 2025 (including the GDPR, AI Act, and Digital Rights Declaration) reveals that neurodivergence and cognitive accessibility remain largely unaddressed. In response, we propose a model for inclusive AI governance based on five pillars: participatory co-design, ethical oversight committees, dynamic regulation, ethical education in cognitive diversity, and representative data. We argue that only by integrating neurodivergence as a structural axis in design, regulation, and evaluation can AI systems become truly fair and just, ensuring technological progress benefits all members of society and preventing the perpetuation of algorithmic injustice.

Keywords: Artificial Intelligence neurodivergence · ethics · inclusion · algorithmic bias · European regulation · accessibility · law

1 Introduction

Artificial intelligence (AI) is no longer a tool of the future: it is an active part of the present. It is integrated into everyday processes such as booking accommodation, receiving medical assistance, or participating in automated job interviews. It also participates in institutional decisions that were previously under human control: access to employment, public management, social services, or justice. However, its technical development has advanced much faster than its ethical framework. One of the major absences along this path has been neurodivergence.

When we talk about neurodivergence, we refer to brain functioning profiles that do not follow the standard pattern expected by current systems: autism, ADHD, dyslexia, Tourette syndrome, among others. These differences are not errors in the human system, but legitimate expressions of how our species functions. But institutional and technological environments, especially those that use AI, are not designed for this diversity. They are designed from what is considered “normal,” and anything that doesn’t fit the definition of ‘normal’ is left out.

This has serious consequences since AI systems are not neutral. They make decisions that affect real lives, and, if they do not recognize or understand cognitive diversity, they reproduce and exacerbate inequalities. A neurodivergent person may be excluded from a selection process because their way of communicating does not match what the algorithm considers “clear”, they may have difficulties receiving assistance if they do not respond to the standard patterns of an application, and they may be wrongly evaluated if their behavior does not fit into the expected categories.

This article argues that it is not possible to continue building systems without taking all people into account. An ethical AI must include neurodivergence from the outset: in design, in data it is trained on, in regulation, and in evaluation. To develop this argument, we will analyze the current European context, explain why the norms are not sufficient, and present a model based on five pillars: participation, ethics, dynamic regulation, education, and representativity. Only if AI reflects real human diversity can it truly be just.

2 Neurodivergence and Artificial Intelligence

2.1 Definition and Social Context

Neurodivergence is not a disorder, nor an obstacle to be corrected. Neurodivergence includes a series of neurocognitive profiles, such as Autism Spectrum Disorder (ASD), Attention Deficit Hyperactivity Disorder (ADHD), dyslexia, dyspraxia, or Tourette syndrome (Armstrong, 2010; Chapman, 2020). These conditions, far from constituting clinical deficits, represent legitimate functional forms within human neurological plurality (Silberman, 2015). However, most AI systems operate under neuronormative reference models, i.e., configurations based on cognitive patterns considered standard (Singer, 2017), which generates structurally exclusionary biases.

2.2 Documented Exclusion

Various studies have documented how AI systems tend to penalize individuals whose communicative or behavioral patterns do not match the majority data present in training sets (Li & Yang, 2024; Brandsen et al., 2024). This phenomenon not only produces technical errors but also creates structural mechanisms of automated discrimination (Lemke, Bloomfield & Herfurth, 2024; Caliskan, Bryson, and Narayanan, 2017).

In personnel selection, for example, tools such as HireVue or systems based on facial and prosodic analysis can automatically reject candidates whose verbal expressiveness, eye contact, or response rhythm differs from the statistical models employed, without

considering the functional validity of these profiles. These practices, although formally neutral, are in practice discriminatory for neurodivergent people.

This exclusion is also manifested in customer service systems and automated platforms. For example, Airbnb has documented how technical assistance requests submitted by people with direct, emotional, or non-sequential communication styles -frequent characteristics in neurodivergent profiles- are systematically rejected by automated conversational systems that lack mechanisms to interpret communicative diversity. The impossibility of accessing a human operator or providing additional context aggravates the situation, leading to unilateral cancellations and denial of service without legitimate grounds.

Even in environments that use advanced conversational technologies-such as large language models (LLMs) implemented on platforms like ChatGPT-responses have been observed that interrupt, restructure, or invalidate non-conventional forms of reasoning, even when these are valid from an internally coherent logic. Recent autoethnographic research (Liu et al., 2023; Jakesch et al., 2023) has found that these systems tend to reformulate texts when intense argumentative structures, non-linear thinking, or expressive styles with high emotional markers are introduced. These responses are not based on grammatical errors, but on hidden assumptions about what constitutes a “correct” or “efficient” interaction. Ultimately, what is penalized is not the content, but the way of thinking.

This pattern is reproduced in automated educational applications, digital assistants, translation systems, and algorithmic recommendations, where the lack of cognitive representativity in training data generates biased outputs. For example, linguistic correction algorithms may mark as errors literal or direct expressions that are typical of autistic profiles, without offering any adaptive alternative.

2.3 Structural Problem and Ethical Imperative

These cases, far from being exceptional, reflect a structural problem (Pattaro et al., 2005; Sartor, 2020). Current AI not only fails to recognize neurodivergence, but activates exclusion mechanisms through its own operational logic. Technical accessibility without cognitive accessibility is a false solution. If the design of systems does not consider multiple forms of processing, communication, and reasoning, a form of algorithmic injustice is consolidated that penalizes diversity as if it were an error.

The inclusion of neurodivergence in the design, development, validation, and supervision of AI should not be understood as an optional improvement, but as an essential condition for technological equity and the ethical legitimacy of systems (Dotch and Arnold, 2024). Ignoring this dimension means reproducing, in automated form, historical inequalities already documented-and denounced-in other social spheres (Mittelstadt et al., 2016; Floridi & Cowls, 2022).

3 European Legal and Policy Framework

3.1 Overview

Europe has consolidated in 2025 a more rigorous and structured legal framework to regulate the development, deployment, and use of artificial intelligence (European Commission, 2019). The stated objective is to align technological innovation with the European Union's fundamental values: human dignity, equity, social justice, and fundamental rights. However, despite the growing recognition of the need for human-centered AI, current legal frameworks continue to omit a key category: cognitive diversity. Neurodivergence has not yet been integrated as a structural principle or as a dimension of ethical, technical, or legal evaluation.

3.2 General Data Protection Regulation – GDPR (Regulation (EU) 2016/679)

The General Data Protection Regulation (European Parliament & Council, 2016) remains the cornerstone in the protection of personal data and the regulation of automated decisions. Article 22 establishes the right of every person not to be subject to decisions based solely on automated processing, including profiling, that produce significant legal effects. It also requires transparency, accountability, and the possibility of human intervention.

However, the GDPR does not require AI systems to adapt to different cognitive profiles nor does it contemplate reasonable adjustments to ensure cognitive accessibility. Interfaces, algorithmic logics, and automated decision flows continue to be designed for a neurotypical user, which leaves neurodivergent people without effective guarantees against invisible discrimination or algorithmic misinterpretation.

3.3 Directive 2000/78/EC and the Digital Equality Package (2025)

Directive 2000/78/EC (European Parliament & Council, 2020) establishes a general framework for equal treatment in employment. Although it prohibits discrimination on the grounds of disability, the definition of “disability” does not explicitly include neurodivergent profiles, which limits its protective capacity against algorithmic tools that discriminate without explicit intent.

The Digital Equality and Rights Package, currently under discussion, proposes a comprehensive reform of this directive to include algorithmic discrimination and broaden the concept of disability to neurofunctional conditions. However, until this reform is approved, effective legal protection against digital exclusion remains very limited.

3.4 Regulation (EU) 2024/1689 on Artificial Intelligence – AI Act

The Artificial Intelligence Act (European Parliament & Council, 2024), approved on March 13, 2024, is the world's first binding legal framework on artificial intelligence. It classifies systems by risk levels (minimal, limited, high, prohibited), requires transparency, human oversight, technical documentation, and establishes significant penalties for non-compliance.

Despite its ambition, the AI Act does not mention neurodivergence in any of its articles, annexes, or requirements. It does not require specific measures to include diverse cognitive profiles in the design, testing, or auditing of systems. Accessibility is addressed from a technical perspective, but not from the perspective of neurological diversity, perpetuating the legal invisibility of millions of people.

3.5 European Declaration of Digital Rights and Principles (2022–2030)

This political declaration establishes the value framework for Europe’s digital transformation. It recognizes the right to inclusive, safe, and accessible digital environments for all people, and promotes equity in access to technologies.

Although it is not legally binding, it is a key instrument for demanding normative coherence. In its 2024 update, it mentions for the first time the need for “digitally accessible environments for all human diversity,” providing institutional legitimacy to demand the operational integration of neurodivergence in the design of digital policies.

3.6 European AI Strategy and Horizon Europe Programs (2024–2027)

The deployment of the European AI strategy combines regulation, investment, and research. It funds centers of excellence, promotes cross-border alliances, and encourages the training of ethical and technical profiles. In 2025, the Horizon Europe and Digital Europe programs support projects in algorithmic governance, explainability, and automated justice.

However, neurodivergence continues to be unrepresented as a specific category in calls, technical standards, or ethical audit mechanisms. No cognitive equity indicators or inclusive design guidelines for developers have been created. This absence consolidates a structural gap between political discourse and technical implementation.

3.7 Conclusion of the Legal Analysis

The European Union has taken decisive steps to build a regulated, ethical, and people-centered AI framework. But neurodivergence remains off the regulatory radar. Current regulations do not consider this diversity as a legal, technical, or operational dimension. This prevents the detection, prevention, and correction of neuronormative biases that systematically occur in selection, evaluation, diagnosis, or rights access systems.

If Europe aspires to lead a just Artificial Intelligence model, it needs to integrate cognitive diversity into its regulatory frameworks, not as a footnote, but as a structural axis. Only then will it be possible to ensure that AI not only works well, but works for everyone.

4 Proposed Model: Five Pillars for Inclusive AI

An ethical and truly just Artificial Intelligence cannot be built from an exclusively technical approach nor from regulatory frameworks that omit cognitive diversity. To move towards inclusive governance, it is necessary to propose operational models that integrate, transversally, neurodivergent people not only as recipients, but as active subjects of the technological development process.

This work proposes a model articulated in five fundamental pillars. Each responds to a structural gap detected in the legal and ethical analysis carried out above. These are not voluntary recommendations, but minimum conditions to ensure that AI systems respect the principle of cognitive equity, representative participation, and effective human supervision. Table 1 provides a summary of these basic elements for an inclusive AI.

Table 1. Summary of the Five Pillars for Inclusive AI

Pillar	Description
Participatory Co-design	Effective inclusion of neurodivergent individuals in the early stages of system design and testing. This means active participation in defining objectives, interfaces, interpretation criteria, and validation of results. User experience cannot be defined without real users with diverse profiles. Including neurodivergent people from the beginning not only improves technical quality but also ensures ethical legitimacy
Ethical Oversight Committees	The presence of independent ethical committees, with representation from neurodivergent profiles, as well as professionals in ethics, law, medicine, and technology, must be a requirement for high-impact systems. These committees must be able to assess specific risks, demand modifications, and guarantee effective complaint mechanisms in case of harm resulting from system operation
Dynamic Regulation	Regulations must incorporate mechanisms for periodic review, algorithmic impact audits, and cognitive bias assessment. The static approach of technological legislation does not allow adaptation to changes in practice and the evolution of knowledge. It is essential to include specific cognitive equity indicators and corrective mechanisms when exclusion patterns are detected
Ethical Education in Cognitive Diversity	Training those who design, regulate, and apply artificial intelligence is a structural measure. Inclusive ethics should be part of university and professional training, incorporating content on neurodivergence, epistemic biases, algorithmic justice, and digital rights. Without a deep understanding of human diversity, it is not possible to ensure fair AI development
Representative Data	Systems trained with homogeneous data reproduce limited views of human behavior. To avoid neuronormative biases, it is necessary to ensure that data reflects the plurality of cognitive expressions. This includes language patterns, non-verbal communication, reasoning styles, and diverse responses. Representativity is not an ideal: it is a technical and ethical requirement

4.1 Participatory Co-design

The effective inclusion of neurodivergent people in the initial phases of system design and testing is essential. This implies their active participation in defining objectives, interfaces, interpretation criteria, and validation of results. User experience cannot be defined without real users with diverse profiles. Including neurodivergent people from the outset not only improves technical quality, but guarantees ethical legitimacy (Shinohara & Wobbrock, 2011).

4.2 Ethical Oversight Committees

The presence of independent ethical committees, with representation from neurodivergent profiles, professionals in ethics, law, medicine, and technology, should be a

requirement in high-impact systems. These committees must have the capacity to assess specific risks, demand modifications, and guarantee effective complaint mechanisms in case of harm resulting from system operation.

4.3 Dynamic Regulation

Norms must incorporate mechanisms for periodic review, algorithmic impact audits, and cognitive bias assessment. The static approach of technological legislation does not allow adaptation to changes in practice and the evolution of knowledge. It is essential to include specific cognitive equity indicators, and corrective mechanisms when exclusion patterns are detected.

4.4 Ethical Education in Cognitive Diversity

Training those who design, regulate, and apply artificial intelligence is a structural measure. Inclusive ethics should be part of university and professional training, incorporating content on neurodivergence, epistemic biases, algorithmic justice, and digital rights. Without a deep understanding of human diversity, it is not possible to ensure fair AI development.

4.5 Representative Data

Systems trained with homogeneous data reproduce limited views of human behavior. To avoid neuronormative biases, it is necessary to ensure that data reflect the plurality of cognitive expressions. This includes language patterns, non-verbal communication, reasoning styles, and diverse responses. Representativity is not an ideal: it is a technical and ethical requirement.

5 Conclusion

The expansion of artificial intelligence (AI) into critical decision-making processes is reshaping access to rights, services, and participation in society. However, this transformation is unfolding within cognitive, legal, and technical frameworks that largely overlook neurodivergence. As a result, individuals with non-normative cognitive profiles continue to be structurally excluded from AI systems that are presumed to be neutral and universal.

Current European regulations, including the General Data Protection Regulation (European Parliament & Council, 2016), Directive 2000/78/EC (European Parliament & Council, 2020), and the AI Act (European Parliament & Council, 2024), have contributed to building a more transparent and accountable AI ecosystem. Yet, none of these frameworks explicitly address neurodivergence as a category of protection or design requirement. This omission enables algorithmic systems to reproduce and amplify cognitive exclusion without ethical or legal safeguards.

This paper has demonstrated that algorithmic discrimination against neurodivergent individuals is not limited to employment or education. It occurs systematically in digital

platforms, automated customer services, language models, and decision-making tools that normalize specific patterns of interaction and penalize cognitive diversity. These are not isolated design flaws, but expressions of an epistemic and structural bias embedded in AI development.

In response, we have proposed a governance model grounded in five structural pillars: participatory co-design, ethical oversight with neurodivergent representation, dynamic regulation with cognitive equity indicators, ethics education in cognitive diversity, and the use of representative datasets. These are not supplementary recommendations but necessary conditions for ensuring ethical and legitimate AI systems.

Cognitive equity must no longer remain an abstract ethical aspiration; it must be a binding normative requirement. Recognizing neurodivergent profiles as integral to human diversity is essential for building AI that reflects the society it claims to serve. Failing to do so risks consolidating a form of structural algorithmic injustice that will perpetuate historical inequalities in a technologically automated form.

Only a deliberate, inclusive redesign of AI systems—aligned with principles of cognitive justice and digital rights—will enable the development of technologies that are truly democratic, transparent, and equitable.

Disclosure of Interests. The authors declare no competing interests relevant to the content of this article, which has not been influenced by any external funding or institutional conflict.

References

- Armstrong, T.: *Neurodiversity: discovering the extraordinary gifts of autism, ADHD, dyslexia, and other brain differences*. Da Capo Press, Cambridge (2010)
- Brandsen, S., Chandrasekhar, T., Franz, L., Grapel, J., Dawson, G., Carlson, D.: Prevalence of bias against neurodivergence-related terms in artificial intelligence language models. *Autism Res.* **17**(2), 234–248 (2024). <https://doi.org/10.1002/aur.3094>
- Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. *Science* **356**(6334), 183–186 (2017). <https://doi.org/10.1126/science.aal4230>
- Chapman, R.: Defining neurodiversity for research and practice. In: Bertilsdotter Rosqvist, H., Chown, N., Stenning, A. (eds): *Neurodiversity Studies: A New Critical Paradigm*, pp. 218–220. Routledge (2020)
- Dotch, E., Arnold, V.: Inclusive practices for child-centered AI design and testing. arXiv preprint [arXiv:2404.05920](https://arxiv.org/abs/2404.05920) (2024). <https://doi.org/10.48550/arXiv.2404.05920>
- European Commission: ethics guidelines for trustworthy AI. Publications Office of the European Union (2019). <https://data.europa.eu/doi/https://doi.org/10.2759/346720>
- European parliament & council: directive 2000/78/EC of the European parliament and of the council establishing a general framework for equal treatment in employment and occupation. Official J. Eur. Union (2000). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32000L0078>
- European parliament & council: regulation (EU) 2016/679 of the European parliament and of the council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official J. Eur. Union (2016). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679>

- European parliament & council: regulation (EU) 2024/1689 of the European parliament and of the council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending various regulations and directives (Artificial Intelligence Act). Official J. Eur. Union (2024). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689>
- Eusebio, E.: Neurodiversity. In: Kreutzer, J., DeLuca, J., Caplan, B. (eds) *Encyclopedia of Clinical Neuropsychology*. Springer (2017). https://doi.org/10.1007/978-3-319-56782-2_9153-1
- Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. In: S. Carta (ed.): *Machine Learning and the City: Applications in Architecture and Urban Design*, pp. 535–545. John Wiley & Sons Ltd. (2022). <https://doi.org/10.1002/9781119815075.ch45>
- Horvat, T., Horvat, A.: Artificial intelligence and neurodiversity: towards inclusive systems. *Sci. et Fides*, **13**(1), 215–232 (2025). <https://doi.org/10.12775/SetF.2025.011>
- Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., Naaman, M.: Co-writing with opinionated language models affects users' views. In: *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–15 (2023). <https://doi.org/10.1145/3544548.3581196>
- Lemke, C., Bloomfield, M., Herfurth, F. N.: Unveiling neuronormative bias: exposing an unconscious bias in generative AI and their impact on neurodivergent people. *NeuroConverse J.* (2024)
- Li, L. Z., Yang, P.: Artificial Intelligence for neurodiversity. *BMJ* (2024). <https://doi.org/10.1136/bmj.q1384>
- Liu, Y., et al.: ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models. *arXiv preprint arXiv:2304.07666*. (2023). https://ui.adsabs.harvard.edu/link_gateway/2023arXiv230407666L/https://doi.org/10.48550/arXiv.2304.07666
- Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: Mapping the debate. *Big Data Soc.* **3**(2), 1–21 (2016). <https://doi.org/10.1177/2053951716679679>
- Pattaro, E., Rottleuthner, H., Shiner, R.A., Peczenik, A., Sartor, G.: Collective cognition and dialogues. In Roversi, C. (ed.): *A treatise of legal philosophy and general jurisprudence* (Vol. 1), pp. 1287–1313. Springer. (2005). https://doi.org/10.1007/1-4020-3505-5_50
- Sartor, G.: Artificial Intelligence and human rights: between law and ethics. *Maastricht J. Eur. Comp. Law* **27**(6), 705–719 (2020). <https://doi.org/10.1177/1023263X20981566>
- Shinohara, K., Wobbrock, J.O.: In the shadow of misperception: assistive technology use and social interactions. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 705–714. (2011). <https://doi.org/10.1145/1978942.1979044>
- Silberman, S.: *NeuroTribes: The legacy of autism and the future of neurodiversity*. Avery, New York (2015)
- Singer, J.: *Neurodiversity: The birth of an idea*. Judy Singer (2017)



How AI is Reshaping Creativity: DeepSeek vs ChatGPT Plus in LEGO® SERIOUS PLAY®

António Almeida¹(✉), Denise Meyerson², and Inês Almeida^{3,4,5}

¹ Instituto Superior de Ciências Sociais e Políticas da Universidade de Lisboa, Lisbon, Portugal
a_jose_almeida@sapo.pt

² University of the Witwatersrand, Johannesburg, South Africa

³ Autónoma Techlab, Universidade Autónoma de Lisboa, Lisbon, Portugal

⁴ Instituto de Telecomunicações, Aveiro, Portugal

⁵ FCT, Nova University of Lisbon, Caparica, Portugal

Abstract. This study investigates how generative artificial intelligence (AI) models—specifically ChatGPT Plus (with and without the Projects feature) and DeepSeek—can support the design of LEGO® SERIOUS PLAY® session plans. Through a systematic reflexive content analysis methodology, we evaluate the originality, structure, and methodological fidelity of AI-generated LSP session plans across two distinct thematic areas: Creative Leadership and Building Team Trust. Our multi-layered evaluation framework includes content analysis, session structure mapping, and a temporal triangulation protocol that compensates for the limitations of single-researcher evaluation. Findings reveal that ChatGPT Plus with Projects exhibits superior consistency and alignment with LSP principles (scoring 13.7/15 on our composite index), while DeepSeek demonstrates strengths in information synthesis but limited session structuring (8.9/15). Comparative analysis of AI-generated facilitator prompts reveals significant qualitative differences, with ChatGPT Plus with Projects producing questions with greater metaphorical depth and nuance, creating space for participants to explore complex dimensions of the themes. The study contributes to the fields of creativity, participatory facilitation, and AI by providing an evidence-based framework for assessing AI support in creative methodologies and offers practical guidelines for facilitators seeking to leverage AI as a collaborative partner in session design. Our prompt engineering protocol, detailed in this paper, provides a replicable template for facilitators and researchers exploring AI-human collaboration in experiential learning design.

Keywords: Creativity · Artificial Intelligence · LEGO® SERIOUS PLAY® · Generative AI · ChatGPT Plus; DeepSeek

1 Introduction

Creativity can be described as the ability to generate new ideas, concepts, or solutions that are original and useful within a given context [1]. LEGO® SERIOUS PLAY® is a methodology that uses LEGO® bricks as a facilitating tool to enhance systemic creativity

and problem-solving [2]. Based on theories of play, constructivism, and constructionism, this approach encourages participants, during group sessions, to build models representing professional and/or personal scenarios, thereby stimulating their creative thinking [3].

The playful environment of LEGO® SERIOUS PLAY® creates a relaxed and familiar space, reducing psychological barriers and encouraging experimentation, allowing all participants to explore new perspectives without limitations, significantly expanding their creative potential [4]. In each session, participants receive a set of LEGO® bricks, and the facilitator presents a challenge in the form of a question, which must be answered by building a model [5]. Each participant has a limited amount of time to construct their model and, at the end of this period, is encouraged to explain their construction through an individual narrative. The act of building with hands and interacting with the model allows participants to externalize their mental models, encouraging them to express their ideas intuitively and spontaneously [6]. Furthermore, the verbalization process activates the emotional and associative side of the brain, promoting creative and authentic solutions [7]. At the end of each session, participants collaborate in building a collective model, where they are encouraged to exchange ideas and integrate different individual perspectives into a shared narrative. The process follows a structured flow in which each individual constructs their vision of a specific topic, providing a diversity of approaches that, when combined, result in a more robust and innovative collective perspective [8].

The LEGO® SERIOUS PLAY® methodology enables participants to externalize and visualize abstract ideas, transforming them into concrete representations through metaphors that help overcome creative blocks [9]. Unlike traditional approaches, where logical analysis precedes creation, LEGO® SERIOUS PLAY® encourages participants to build first and rationalize afterward, fostering a more fluid and innovative thought process [10].

Currently, Artificial Intelligence (AI) plays an increasingly relevant role in facilitating creative processes. In the context of LEGO® SERIOUS PLAY®, AI can enhance creativity when used to support session design, known as the session plan, following the principles of LEGO® SERIOUS PLAY® [11]. AI can generate ideas, structure scripts, create facilitating questions, and adapt the process according to participants objectives and profiles. Each LEGO® SERIOUS PLAY® session plan comprises different phases, including briefing, warm-up, individual activities, collective dynamics, and debriefing [11].

One key challenge when integrating AI into the LEGO® SERIOUS PLAY® methodology is ensuring that generative AI models truly understand and adhere to the structured facilitation process. A potential avenue for exploration is the development of a dedicated GPT model specifically trained to generate session plans aligned with the official LEGO® SERIOUS PLAY® methodology. Such a model could be fine-tuned with structured input, incorporating elements such as client context, desired outcomes, participant profiles, and time constraints. This structured approach would enhance the reliability and consistency of AI-generated session plans, mitigating the common issue where general AI models sometimes deviate from the methodology and instead generate loosely structured activities that resemble free play rather than a guided facilitation process [12].

This study analyses the role of two generative AI tools, ChatGPT Plus and DeepSeek, in generating session plans and facilitating questions for LEGO® SERIOUS PLAY® to enhance creativity. ChatGPT, developed by OpenAI, uses the Generative Pre-trained Transformer (GPT) architecture and stands out for its ability to process natural language, adapt to different contexts, and generate coherent and relevant responses. ChatGPT Plus, the premium version of OpenAI chatbot, provides access to GPT-4, which offers improved reasoning capabilities, longer context windows, and more accurate responses compared to the free version, which is limited to GPT-3.5. DeepSeek [13] is an emerging open-source alternative that allows anyone to access, modify, and distribute the software's source code for free. This generative AI tool differentiates itself by focusing on information retrieval and synthesis, offering an approach more oriented toward knowledge extraction and response optimization [14]. Both models can be used to stimulate creativity and structure session plans in the LEGO® SERIOUS PLAY® context, expanding possibilities for innovation, collaboration, and communication. These two generative AI tools offer distinct features that may benefit different user needs [12].

To improve the validity of AI-generated session plans, this study also explores the use of standardized session templates, outlining facilitator objectives, expected outcomes, participant demographics, and available time constraints. By providing AI models with structured input, the study aims to compare the effectiveness of different AI generated session plans more reliably, ensuring that results are evaluated within a standardized framework rather than in an ad hoc manner [11].

This study contributes theoretically to the intersection of creativity, participatory methodologies, and artificial intelligence by exploring how generative AI tools can support and transform the facilitation of LEGO® SERIOUS PLAY® session plans [2]. The research is based on a comparative study between ChatGPT Plus and DeepSeek in generating session plans for LEGO® SERIOUS PLAY® that enhance creativity. The study employs a qualitative and experimental approach, where multiple AI-generated session plans are analyzed and compared in terms of originality, adequacy, and applicability [5]. Additionally, the session plans and their respective facilitating questions created by AI models are examined and compared with those developed by human facilitators [3]. The expected results include evaluating AI's ability to structure effective LEGO® SERIOUS PLAY® sessions, identifying patterns and limitations in AI-generated session plans, comparing the diversity between automatically generated and manually designed questions, and exploring AI impact on facilitation dynamics and participant autonomy [6].

This paper explores this frontier by comparing how three AI configurations—ChatGPT Plus (GPT-4), ChatGPT Plus with Projects, and DeepSeek—perform in generating LEGO® SERIOUS PLAY® session plans. Rather than implementing the plans in real sessions, we take a methodological approach, evaluating the quality of the generated outputs through a rigorous comparative lens. The central question is whether AI can support or even transform the way facilitators approach LEGO® SERIOUS PLAY® design. This study aims to:

1. Evaluate the capability of different AI tools to generate methodologically sound LEGO® SERIOUS PLAY® session plans

2. Compare the quality of AI-generated plans across multiple dimensions including structure, creativity, and fidelity to LEGO® SERIOUS PLAY® principles
3. Identify patterns of strengths and weaknesses across different AI architectures
4. Develop a replicable prompt engineering protocol for LEGO® SERIOUS PLAY® facilitators seeking to use AI as a design partner

Building on this foundation, this study further investigates which generative AI tool: ChatGPT Plus or DeepSeek best meets the needs of LEGO® SERIOUS PLAY® facilitators, considering factors such as generating creative questions, structuring sessions, and supporting the facilitation process. The comparative analysis explores how each model responds to facilitators' requirements in terms of flexibility, relevance of suggestions, and practical applicability. Although both tools have the potential to optimize different aspects of LEGO® SERIOUS PLAY®, this study aims to identify which one offers more effective support and aligns better with session dynamics, providing an empirical basis for integrating AI into the facilitation of creative processes [13]. The significance of this research extends beyond LEGO® SERIOUS PLAY® to the broader field of creative facilitation, providing insights into how AI might complement human expertise in designing participatory experiences. As Mollick and Mollick suggest, AI has the potential to serve as a "thought partner" rather than a replacement for human creativity, potentially enhancing rather than diminishing the facilitator's role.

2 Methodology

2.1 Research Design

The research was structured in four sequential phases, each employing specific techniques to ensure methodological rigor:

1. **Generation Phase:** Creation of session plans using standardized prompts across all AI tools
2. **Primary Evaluation Phase:** Initial assessment using our multi-layered rubric
3. **Temporal Re-evaluation Phase:** Reassessment after 14 days to verify analytical consistency
4. **Final Comparative Analysis:** Cross-case examination to identify patterns and variations

This phased approach, combined with rigorous documentation of reflexive processes, ensured that methodological limitations of single-researcher evaluation were systematically addressed and compensated for through temporal triangulation.

2.2 Tools and Data Collection

Three generative AI tools were selected based on their accessibility, prominence, and technical characteristics, ChatGPT Plus (GPT-4), ChatGPT Plus with Projects and DeepSeek.

All interactions with the AI tools were conducted between February 16 and March 16, 2025, using the most current versions available at that time. For DeepSeek and

ChatGPT Plus, dedicated chats were used, while for ChatGPT Plus with Projects, we created a dedicated project titled “LEGO® SERIOUS PLAY® Workshop Designer” that included additional context about LEGO® SERIOUS PLAY® theoretical foundations from official training materials.

To ensure consistency and valid comparison, an identical prompt template was used across all three AI tools. The prompt used was:

Role: You are an expert LEGO SERIOUS PLAY facilitator with certification and 5+ years of experience designing workshop experiences.

Task: Create a detailed 90-minute LEGO SERIOUS PLAY workshop plan on the theme of <Theme>.

Audience: The session will involve 12 participants who are mid-level managers in a technology company.

Goals: The primary goals of this session are to:

1. Help participants explore their understanding of [THEME]
2. Identify personal and organizational barriers related to [THEME]
3. Co-create solutions that address these barriers
4. Develop actionable next steps for implementation

Methodology: Your plan must follow LEGO® SERIOUS PLAY® methodology including:

- A skills building warm-up phase (5-15 minutes)
 - Individual model building before sharing (everyone builds, everyone shares)
 - Clear, open-ended questions that encourage metaphorical thinking
 - A sequence that builds from individual to shared models
 - Time allocations for each phase
-
- Adherence to the principle that "the builder owns the meaning"
 - Story-first, questions-later approach to model sharing

Format: Structure your plan with clear headings, timing, instructions, and facilitator notes.

Output: Provide a complete, ready-to-use session plan that a certified facilitator could implement without additional preparation.

This prompt was systematically applied for each thematic area (Creative Leadership and Building Team Trust) across all three AI tools, ensuring methodological consistency while allowing for valid comparison of outputs.

2.2.1 Corpus

Each tool generated two session plans based on the two different themes. In total, nine session plans were produced by AI. For comparative purposes, we included a reference plan authored by a certified LEGO® SERIOUS PLAY® facilitator with over 5 years of experience, using the same thematic constraints. The data collection process involved recording all interactions with the AI tools, including any requests for clarification. It also included cataloging the generated plans, analyzing their structure, and identifying any unusual or innovative elements. Additionally, structure maps were created to visualize the flow and components of each plan. Finally, similarities and differences between plans addressing the same theme were documented.

2.3 Evaluation Framework

We developed a multi-layered assessment approach that triangulates different analytical perspectives and incorporates structured reflexivity based on Structured Content Analysis, Content Analysis and Methodological Fidelity Assessment.

2.3.1 Structured Content Analysis

We performed a systematic mapping of each session plan against the canonical LEGO® SERIOUS PLAY® structure, identifying the presence, sequencing, and time allocation for each required phase:

- Introduction and Context Setting
- Skills Building (Warm-up)
- Individual Building Challenge
- Sharing and Storytelling
- Collective Model Building
- Reflection and Meaning-making
- Closure and Action Planning

This mapping produced a structural completeness score (0–5) and identified any methodological omissions or innovations.

2.3.2 Content Analysis

We conducted a detailed content analysis of the building prompts and facilitation instructions, examining:

- **Question Quality:** Openness, metaphorical potential, and alignment with thematic goals
- **Language Patterns:** Instructional clarity, emotional resonance, and inclusivity
- **Facilitation Guidance:** Appropriateness of suggested interventions and debriefing questions

This analysis yielded a creativity score (0–5) and insights into the linguistic patterns of each AI model.

2.3.3 Methodological Fidelity Assessment

Using a checklist derived from official LEGO® SERIOUS PLAY® training materials, we evaluated each plan's adherence to core LEGO® SERIOUS PLAY® principles:

- Build First, Talk Later
- Everyone Builds, Everyone Shares
- The Builder Owns the Meaning
- Questions Before Answers

This assessment produced a fidelity score (0–5) and identified any principles that were consistently misunderstood or omitted by the AI models.

2.4 Reflexivity Practices and Reliability Measures

To enhance the reliability of our evaluation despite the constraint of limited evaluators, we implemented several methodological safeguards centered around structured reflexivity:

- **Structured Reflexive Journal:** Throughout the analytical process, we maintained a structured reflexive journal documenting initial impressions before systematic coding, challenges in applying evaluation criteria, emergent questions during analysis, reflections on potential personal biases, and decision-making rationales.
- **Temporal Triangulation:** The evaluator assessed each plan twice, with a 14-day interval between evaluations, to check for consistency. This temporal separation created cognitive distance between assessments, allowing inconsistencies or shifts in interpretation to surface naturally.
- **Structured Coding Protocol:** We developed explicit coding rules for each evaluation dimension, ensuring consistent application of criteria.
- **External Face Validation:** We shared selected plans with a certified LEGO® SERIOUS PLAY® facilitator external to the research team for feedback on face validity.

3 Results

Our evaluation revealed significant differences in how each AI tool approached the generation of LEGO® SERIOUS PLAY® session plans. Table 1 presents the composite scores across our three assessment dimensions.

ChatGPT Plus with Projects consistently outperformed the other AI tools, with composite scores approaching those of the human facilitator reference plan. DeepSeek demonstrated particular weaknesses in methodological fidelity, often diverging from core LEGO® SERIOUS PLAY® principles.

The structural completeness of session plans varied significantly between tools. ChatGPT Plus with Projects closely mirrored the human facilitator balance across phases, particularly in allocating sufficient time for collective model building and reflection. DeepSeek consistently overallocated time to introductory content while underallocating time for the critical collective model building phase.

The nature of building prompts and facilitation questions proved to be a key differentiator between tools. Table 2 presents a qualitative comparison of representative building prompts from each AI tool.

Table 1. Composite Evaluation Scores by AI Tool and Theme

AI Tool	Theme	Structure Score (0–5)	Creativity Score (0–5)	Fidelity Score (0–5)	Composite Score (0–15)
ChatGPT Plus	Creative Leadership	4.3	3.8	4.1	12.2
ChatGPT Plus	Building Team Trust	4.1	3.9	4.0	12.0
ChatGPT Plus with Projects	Creative Leadership	4.7	4.5	4.6	13.8
ChatGPT Plus with Projects	Building Team Trust	4.6	4.3	4.7	13.6
DeepSeek	Creative Leadership	3.2	3.8	2.0	9.0
DeepSeek	Building Team Trust	3.1	3.7	2.1	8.9
Human Facilitator (Reference)	Creative Leadership	4.9	4.8	5.0	14.7

Table 2. Representative Building Prompt Comparison (Building Team Trust Theme)

AI Tool	Building Prompt	Analysis
ChatGPT Plus	“Using your LEGO bricks, build a model that represents what trust means to you in your team.”	Basic metaphorical prompt, somewhat generic
ChatGPT Plus with Projects	“Using your LEGO bricks, construct a model that represents what trust looks like when it’s thriving in your team. Consider both visible and invisible elements that contribute to this trust.”	Rich metaphorical potential, layered complexity, invites deeper reflection
DeepSeek	“Build a model showing team trust.”	Directive, closed, limited metaphorical potential

ChatGPT Plus with Projects consistently generated questions with greater metaphorical depth and nuance, creating space for participants to explore complex dimensions of the themes. DeepSeek’s questions tended toward the literal and directive, limiting creative exploration.

Adherence to core LEGO® SERIOUS PLAY® principles varied significantly across tools. ChatGPT Plus with Projects demonstrated strong alignment with all four key principles, while DeepSeek showed consistent misunderstandings, particularly around “The Builder Owns the Meaning” and “Build First, Talk Later” principles.

A pattern emerged where DeepSeek frequently included directive interpretations of participant models, violating the principle that meaning belongs to the builder. For example, in the Building Team Trust session plan, DeepSeek included facilitator notes such as: “If participants build walls or barriers, highlight how these represent communication challenges that need addressing.”

This contrasts sharply with ChatGPT Plus with Projects, which consistently included guidance like: “Listen carefully to each participant’s story without interpretation. Ask clarifying questions only after they have fully shared the meaning of their model.”

Each AI tool demonstrated different innovative approaches to session design. ChatGPT Plus with Projects uniquely incorporated layered building challenges that built upon previous models, creating narrative continuity throughout the session. For example, in the Creative Leadership theme, it suggested: “Now add to your leadership model something that represents how your leadership approach changes under pressure.”

DeepSeek, despite methodological limitations, excelled at synthesizing relevant theoretical frameworks into session introductions, drawing connections to established leadership and team dynamics theories. However, this sometimes came at the expense of actual building time.

4 Conclusion

This study provides empirical evidence that generative AI tools can support the design of methodologically sound LEGO® SERIOUS PLAY® session plans, though with varying degrees of effectiveness. ChatGPT Plus with Projects emerged as the superior tool, demonstrating strong alignment with LEGO® SERIOUS PLAY® principles and generating session plans that approached the quality of those created by experienced human facilitators.

Our research found that ChatGPT Plus with Projects excelled in structural coherence and methodological fidelity, while DeepSeek demonstrated strengths in theoretical synthesis but struggled with core methodological principles. The addition of specialized project context in ChatGPT Plus with Projects substantially improved output quality, suggesting that AI tools require specific methodological knowledge to generate effective session plans. All AI tools occasionally struggled with optimization and sometimes overstructured the sharing process, potentially limiting emergent conversation. Our standardized prompt template proved effective in guiding AI output toward methodologically sound session plans, highlighting the importance of clear constraint specification.

These findings have significant implications for LEGO® SERIOUS PLAY® practitioners. Rather than replacing human expertise, AI tools like ChatGPT Plus with Projects can serve as valuable thought partners in the session design process, offering creative alternatives and structural templates that facilitators can refine. The most effective application may be a collaborative approach where AI generates initial session structures that human facilitators then customize based on specific client needs and contextual factors.

Limitations of this study include the single-researcher evaluation approach (mitigated through temporal triangulation) and the focus on plan quality rather than implementation outcomes. Future research should explore how AI-generated plans perform in actual workshop settings and investigate how facilitators might most effectively collaborate with AI tools during the design process.

In conclusion, while AI tools cannot yet fully replace human expertise in LEGO® SERIOUS PLAY® facilitation, they demonstrate considerable potential as design partners, particularly when provided with appropriate methodological context. ChatGPT Plus with Projects represents the current state-of-the-art for this application, offering facilitators a valuable resource for creative session design.

References

1. Meyerson, D.: Better, best, brilliant: The essential guide for trainers and facilitators. Post Pre-Press Group (2013)
2. Benesova, N.: LEGO® Serious Play® in management education. *Cogent Educ.* **10**(1), 2262284 (2023)
3. Roos, J., Victor, B.: How it all began: the origins of LEGO® Serious Play®. *Int. J. Manage. Appl. Res.* **5**(4), 326–343 (2018)
4. Zenk, L., Primus, D.J., Sonnenburg, S.: Alone but together: flow experience and its impact on creative output in LEGO® Serious Play®. *Eur. J. Innov. Manage.* **25**(6), 340–364 (2022)
5. Kriszan, A., Nienaber, B.: Researching playfully? Assessing the applicability of LEGO® Serious Play® for researching vulnerable groups. *Societies* **14**(15), 1–11 (2024)
6. Zarndt, F.: Unlocking student creativity with LEGO® Serious Play: a case study from the graduate marketing classroom. *Mark. Educ. Rev.* **34**(2), 105–119 (2024)
7. Nerantzi, C., James, A.: LEGO® for university learning: online, offline and elsewhere (2022)
8. Meyerson, D., Smith, J.L., Walling, S.J.: Strategic play: The creative’s facilitator’s guide. What the Duck! (2017)
9. Tawalbeh, M., Riedel, R., Dempsey, M., Emanuel, C.: LEGO® Serious Play® as a business innovation enabler. Technische Universität Chemnitz (2018)
10. Wheeler, A.: LEGO® Serious Play® and higher education: encouraging creative learning in the academic library. *Insights* **36**(8), 1–8 (2023)
11. Pérez, J., Castro, M., López, G.: Serious games and AI: challenges and opportunities for computational social science. *IEEE Access* **11**, 62051–62061 (2023)
12. Mollick, E.R., Mollick, L.: Instructors as innovators: a future-focused approach to new AI learning opportunities, with prompts. *Wharton Sch. Res. Pap.* (2024)
13. Wang, Y., Pan, Y., Yan, M., Su, Z., Luan, T.H.: A survey on ChatGPT: AI-generated contents, challenges, and solutions. *IEEE Xplore* (2023)
14. Guo, D., et al.: Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning (2025). arXiv preprint [arXiv:2501.12948](https://arxiv.org/abs/2501.12948)



Sovereignty, Surveillance, and the Cloud: Geopolitical and Ethical Issues of Global Cloud Computing

Ala Ali Almahameed¹ , Jorge Pelegrín-Borondo² ,
Juan Luis López-Galiacho Perona³ , and Mario Arias-Oliva^{1,4}  

¹ University Rovira i Virgili, Tarragona, Spain
mario.arias@ucm.es

² University of La Rioja, Logroño, Spain

³ Rey Juan Carlos University, Madrid, Spain

⁴ Complutense University of Madrid, Madrid, Spain

Abstract. This narrative review examines the growing geopolitical and ethical complexities surrounding cloud computing, driven by its increasing adoption across sectors. It offers a comprehensive analysis of how national policies, data sovereignty concerns, and ethical considerations shape global cloud governance. The study analyzes the approaches adopted by key regions, including the United States, Europe, the Middle East, and China, highlighting their strategies for balancing innovation, security, and ethical responsibility. In addition to the regional analysis, the study includes another approach, focusing on challenges within government, banking, telecommunications, healthcare, and military sectors. These critical sectors underscoring the critical need for international cooperation and well-defined regulatory frameworks. The study contributes an interdisciplinary analysis, a comparative study of regional approaches, and insights for various industries. It also identifies future research directions, advocating for a balanced approach prioritizing, among other dimensions, security, privacy, and ethics to ensure the sustainable and responsible development of cloud computing.

Keywords: Cybersecurity · Cloud Computing · Data Sovereignty · Ethics · Privacy

1 Introduction

Cloud computing has revolutionized the way data is stored, processed, and accessed globally. As organizations and governments increasingly rely on cloud infrastructure, geopolitical and ethical concerns have emerged, shaping policies, regulations, and international relations. The intersection of technology, governance, and ethics in cloud computing presents a complex landscape where national security, data sovereignty, privacy, and equitable access are key considerations [1].

Cloud computing has evolved from a niche technology to a fundamental component of digital transformation. Major cloud service providers (CSPs) such as Amazon

Web Services (AWS), Microsoft Azure, Google Cloud, and Alibaba Cloud dominate the market, offering scalable and cost-effective solutions for businesses and governments worldwide. The global cloud computing market is projected to reach \$1.2 trillion by 2028, reflecting its growing economic and technological significance. However, the widespread adoption of cloud computing has raised concerns about data governance, cybersecurity, and ethical implications. Governments and regulatory bodies are increasingly scrutinizing cloud providers to ensure compliance with national laws and international standards. The geopolitical landscape of cloud computing is shaped by competing interests among nations, particularly in the United States, Europe, the Middle East, and China [2].

Geopolitical tensions surrounding cloud computing primarily revolve around data sovereignty, cybersecurity, and technological dominance. Countries seek to protect their digital assets while maintaining control over data generated within their borders. The U.S. Cloud Act, for instance, allows American authorities to access data stored by U.S.-based cloud providers, raising concerns among foreign governments about surveillance and data privacy. Similarly, the European Union has implemented stringent regulations such as the General Data Protection Regulation (GDPR) to safeguard user data and ensure transparency in cloud operations. The EU's GAIA-X initiative aims to establish a sovereign European cloud infrastructure, reducing dependence on American tech giants. These efforts highlight the geopolitical struggle for control over cloud technologies and data governance [3].

In the Middle East, cloud adoption is accelerating, with governments investing in digital infrastructure to drive economic growth. However, cybersecurity threats and geopolitical instability pose challenges to cloud security and data protection. Countries in the region have varying approaches to cloud governance, with some prioritizing state control over digital assets. China, on the other hand, has developed a robust cloud computing industry led by Alibaba Cloud and Tencent Cloud. The Chinese government enforces strict data localization laws, requiring foreign cloud providers to partner with domestic firms to operate within the country. This approach contrasts with Western models advocating open data flows, leading to geopolitical tensions between China and other global powers [2].

Beyond geopolitical concerns, cloud computing raises ethical questions related to privacy, surveillance, and equitable access. The ability of cloud providers to collect and analyze vast amounts of user data has sparked debates on digital rights and corporate responsibility. Ethical considerations in cloud computing include data privacy and user consent, where cloud providers must ensure that users have control over their personal data and understand how it is processed. The GDPR has set a precedent for data protection, emphasizing user rights and transparency. Governments worldwide also leverage cloud technologies for surveillance, intelligence gathering, and cybersecurity operations. The balance between national security and individual privacy remains a contentious issue. Finally, despite its benefits, cloud computing still faces challenges related to the digital divide and unequal access to services. Developing nations face challenges in adopting cloud technologies due to infrastructure limitations and regulatory barriers [4].

As cloud computing becomes increasingly integral to global infrastructure and digital transformation, understanding its geopolitical and ethical implications is critical for policymakers, industry leaders, and researchers. Existing literature often focuses on technical aspects or specific regional perspectives. This review addresses a gap by providing a holistic analysis of the intersection between technology, governance, and ethics in the cloud computing landscape. It is essential to inform strategies that promote secure, equitable, and responsible cloud adoption while addressing the challenges posed by data sovereignty concerns, cybersecurity threats, and differing national interests.

This review aims to provide a comprehensive overview of the geopolitical and ethical landscape of cloud computing; analyze the diverse approaches to cloud governance and regulation adopted by key regions globally, including the United States, Europe, the Middle East, and China; examine sector-specific challenges and considerations related to cloud adoption in government, banking, telecommunications, healthcare, and military sectors; and identify key trends, emerging issues, and potential future directions for cloud computing policy and ethical frameworks.

2 Geopolitical and Ethical Issues by Region

2.1 United States

The United States remains a powerhouse in the global cloud computing industry, primarily through the influence of major providers such as AWS, Microsoft Azure, and Google Cloud. These corporations have a significant impact on both domestic and international policies concerning cloud governance, cybersecurity, and data-sharing agreements across borders. A pivotal aspect of U.S. cloud policy is encapsulated in the controversial U.S. Cloud Act. Enacted in 2018, this legislation allows American law enforcement to demand data from U.S.-based cloud service providers, regardless of where the data physically resides. This has provoked ongoing debates about data sovereignty and raised anxieties among European and Asian policymakers over potential breaches of privacy and increased surveillance [5]. Furthermore, the ongoing U.S.-China technological rivalry has deepened, reflecting broader geopolitical tensions. The U.S. has imposed strict restrictions on Chinese tech firms like Huawei and Alibaba Cloud, citing espionage-related national security risks. These sanctions have considerably reshaped the landscape of global cloud governance, pushing countries to choose sides and sometimes diversify their cloud partnerships to mitigate risks associated with geopolitical instability [6].

Besides these geopolitical dimensions, the ethical landscape within the U.S. cloud computing domain is fraught with complex issues. Privacy rights remain a major concern, especially after revelations about government surveillance using cloud services. Recent studies indicate that there is an increasing demand for transparency and accountability from cloud providers regarding how data is utilized and shared, especially concerning government requests for information [7].

Corporate accountability also emerges as a crucial ethical issue. As cloud providers hold vast amounts of personal and corporate data, their responsibility to prevent misuse and protect data integrity has become paramount. In this context, [8] discussed the

challenges faced by these companies in balancing profit-driven motives with ethical responsibilities, calling for better regulatory frameworks to ensure fair practices.

Another pressing ethical consideration in cloud computing is equitable access to cloud services, particularly in the United States, where the digital divide disproportionately affects marginalized communities. Studies indicate that disparities in digital access are strongly correlated with socioeconomic status, racial identity, and geographic location. Research highlights that low-income households, racial minorities, and rural communities face significant barriers to cloud adoption due to limited infrastructure, affordability issues, and digital literacy gaps [9]. In fact, as digital transformation accelerates, ensuring inclusive access to cloud technologies is crucial for social equity and economic mobility. A systematic review of digital inequity emphasizes the need for policy interventions that address affordability, infrastructure development, and digital literacy programs. Additionally, authors argue that government initiatives and private-sector collaboration are essential to bridging the digital divide and fostering equitable cloud accessibility [10].

Cloud providers are central to intelligence operations, which raises ethical concerns about privacy, surveillance transparency, and civil liberties. Governments worldwide rely on cloud infrastructure for cybersecurity, intelligence analysis, and national security, yet a lack of clear regulatory oversight has fueled debates on accountability. Research highlights opaque agreements between cloud providers and state agencies, emphasizing risks associated with mass surveillance, unrestricted data access, and potential misuse of cloud-stored information. Scholars advocate for stronger oversight mechanisms to ensure ethical governance, transparency, and individual rights protection within cloud-based intelligence frameworks [11].

2.2 Europe

Europe's approach to cloud computing is characterized by a strong emphasis on data sovereignty and privacy, largely driven by the GDPR, which became enforceable in 2018. The GDPR, adopted by the European Union, establishes rigorous data protection standards, ensuring that individuals maintain significant control over their personal data. This regulation has not only reshaped European data governance but has also set a global benchmark, compelling cloud providers worldwide to comply with its provisions to operate within the EU. According to [12], privacy recommendations for cloud services in the public sector further reinforce Europe's commitment to data protection and sovereignty, emphasizing compliance with GDPR to safeguard sensitive information. Research highlights how the GDPR has influenced similar regulatory frameworks in other regions, reinforcing Europe's leadership in data protection and digital sovereignty.

Europe's reliance on U.S.-based cloud providers has raised concerns about data sovereignty and national security, prompting efforts to establish independent cloud infrastructure. The GAIA-X initiative, launched by the European Union, aims to create a federated, secure, and transparent digital ecosystem, reducing dependence on American technology giants while fostering innovation within Europe's digital economy. Recent analyses highlight challenges in implementing GAIA-X, including bureaucratic hurdles

and competition from dominant U.S. cloud providers. Despite these obstacles, GAIA-X remains a strategic effort to enhance European digital sovereignty and strengthen cybersecurity frameworks [13].

Ethical concerns in Europe's cloud computing landscape are primarily centered on transparency, accountability, and user control over personal data. The GDPR has been instrumental in shaping these principles, requiring cloud providers to implement robust data protection measures and establish clear policies on data usage and privacy. According to [14], the GDPR has significantly influenced data governance frameworks, reinforcing Europe's leadership in digital sovereignty and consumer trust. Additionally, European governments are actively promoting cybersecurity initiatives to protect cloud infrastructure from cyber threats and foreign interference. These efforts aim to harmonize security standards across member states and foster international collaboration to combat cybercrime. The European Commission has also emphasized the ethical deployment of artificial intelligence (AI) and cloud computing, advocating for human rights and democratic values in technological governance. As part of this initiative, the EU has outlined principles for fairness, transparency, and accountability in cloud service deployment, ensuring that technology remains a force for innovation while safeguarding citizens' rights.

2.3 Middle East

Cloud computing adoption in the Middle East has experienced rapid growth, driven by both governmental initiatives and private sector investments aimed at fostering digital transformation. Countries across the region recognize the strategic importance of cloud infrastructure to diversify their economies, improve public services, and enhance regional competitiveness. Countries such as the United Arab Emirates (UAE), Saudi Arabia, and Qatar have spearheaded efforts to develop local cloud capabilities and attract multinational cloud providers through favorable regulatory environments. However, the region faces substantial geopolitical instability and cybersecurity vulnerabilities. Ongoing regional conflicts, political tensions, and varying degrees of cooperation among Middle Eastern countries complicate efforts to establish secure, resilient cloud ecosystems. Many nations have adopted differing approaches to data governance, reflecting their strategic priorities. Some prioritize state control over digital infrastructure, exemplified by initiatives such as Saudi Arabia's Vision 2030, which emphasizes local data centers and national cybersecurity strategies. Conversely, other countries may leverage international cloud providers more freely but with stringent regulatory oversight [15].

Ethical concerns in the Middle East center on balancing national security interests with individual privacy rights. Several countries justify their surveillance regimes on security grounds, although they raise concerns about privacy and civil liberties. For example, the extensive use of surveillance technology—often linked to facial recognition and data collection—has prompted debates about privacy infringements. Also, data localization mandates—where countries require data to be stored domestically—present another ethical and geopolitical challenge. While intended to protect national interests, these policies can hinder free data flow, increase operational costs, and limit technological innovation. Gulf Cooperation Council (GCC) countries have actively promoted

data localization policies, ostensibly to secure critical infrastructure and prevent foreign interference, but such policies also risk fragmenting regional digital markets [16].

Cybersecurity risks remain a pressing concern. The Middle East has experienced several notable cyberattacks targeting governmental institutions, financial systems, and critical infrastructure—highlighting vulnerabilities in cloud environments. The region's geopolitical tensions and cyber espionage activities foster an environment where cyber warfare and digital sabotage are increasingly prevalent [17].

Furthermore, investments in cloud infrastructure come with ethical considerations about technology sovereignty, digital literacy, and equitable access. The Gulf nations are heavily investing in smart city projects, AI, and IoT-enabled urban development, reflecting a strategic drive toward digital modernization. However, concerns about data monopolies, user privacy, and the digital divide persist. Although policy initiatives aim to democratize access, disparities between urban centers and rural or marginalized communities can potentially exacerbate inequality, emphasizing the need for inclusive policies [18].

2.4 China

China's rapidly expanding cloud computing industry is primarily driven by domestic technology giants such as Alibaba Cloud, Tencent Cloud, and Huawei Cloud, which have gained significant regional and global market presence. The Chinese government's strategic focus on developing a self-reliant digital infrastructure is supported by comprehensive data sovereignty laws and regulations aimed at maintaining national control over digital assets. One of the central components of this approach is strict data localization laws, which mandate that both domestic and foreign companies store data within China and, in many cases, collaborate with local firms to operate within the country. According to [19], these regulations serve multiple objectives, including enhancing national security, controlling cross-border data flows, and fostering domestic technological innovation. China's approach to cloud governance underscores its commitment to digital sovereignty, ensuring that cloud computing aligns with broader national security and economic development strategies. However, China's approach to internet governance and data regulation reflects a markedly different philosophy from Western open data policies. Instead of promoting free flow of information, China emphasizes "cyber sovereignty"—the principle that states have the right to control and regulate data within their borders. This model often results in rigorous oversight, censorship, and surveillance measures. As a result, foreign cloud providers must partner with local firms to operate in China, adding complexity and increasing compliance costs [20].

Ethical concerns in China are closely linked to its government's surveillance practices and restrictions on digital freedoms. The Chinese Cybersecurity Law, enacted in 2017, mandates that network operators and cloud providers strictly adhere to state regulations, including granting government access to data when required. Critics argue that such policies prioritize national security over individual privacy and corporate accountability, raising concerns about data governance and transparency. China's data policies have significant global implications, particularly affecting international firms seeking to operate within its borders. Multinational technology companies must navigate the complexities of complying with Chinese regulations while upholding their commitments

to user privacy and data protection. These compliance requirements often restrict data handling practices and limit operational autonomy outside China. Beyond corporate concerns, China's regulatory framework reshapes the geopolitical landscape, influencing the deglobalization of internet governance. The strict data localization requirements contribute to a fragmented digital ecosystem, challenging the concept of an open internet promoted by Western nations. This fragmentation carries far-reaching consequences for cross-border commerce, international cooperation, and the development of global standards for cloud security and data privacy [21].

3 Geopolitical and Ethical Issues by Sector

Cloud computing plays a crucial role across various sectors, each facing unique geopolitical and ethical challenges. Reliance on cloud infrastructure for government services, strategic industries (banking, electricity, telecom), healthcare, and military operations raise concerns about data sovereignty, cybersecurity, surveillance, and ethical governance. Below is an in-depth analysis of these sectors, incorporating insights from recent high-ranking journals.

3.1 Government Sector

Governments worldwide are increasingly adopting cloud computing to enhance efficiency, data management, and public service delivery. Its scalability, cost-effectiveness, and improved accessibility make cloud computing attractive for government agencies. However, geopolitical concerns arise due to data sovereignty and foreign cloud provider dependencies. Countries fear that storing sensitive government data on foreign-owned cloud platforms could expose them to surveillance risks or geopolitical leverage. One of the primary concerns is data sovereignty, which refers to a nation's ability to control and protect its data within its jurisdiction. There is concern that foreign cloud providers may be legally obligated to share data with their governments. For instance, the U.S. Cloud Act allows American authorities to access data stored by U.S.-based cloud providers, raising concerns among foreign governments about potential surveillance and data exposure. This has led to efforts by various nations to establish sovereign cloud infrastructures that minimize reliance on foreign technology companies [22].

To mitigate risks associated with foreign cloud providers, several countries have launched initiatives to develop sovereign cloud infrastructures. The European Union's GAIA-X initiative aims to establish a European cloud ecosystem that ensures data sovereignty and reduces dependence on American tech giants. GAIA-X promotes interoperability, security, and transparency, allowing European businesses and governments to retain control over their data while benefiting from cloud computing advancements. Similarly, China has implemented strict data localization policies, requiring foreign companies to store data within the country. This approach ensures that sensitive government and citizen data remain under national jurisdiction, reducing the risk of foreign interference. Other nations, including Russia and India, have also explored data localization strategies to safeguard their digital assets [23].

Beyond geopolitical considerations, ethical concerns surrounding government cloud computing include citizen privacy, government transparency, and the potential misuse of cloud-stored data for surveillance. Governments must balance the benefits of cloud computing with the need to protect citizens' privacy rights. Public trust may be undermined by unauthorized data access or large-scale surveillance programs, raising serious ethical dilemmas. Transparency in government operations is another critical issue. Cloud computing enables efficient data management, but it also raises questions about accountability and oversight. Governments must implement robust policies to ensure that cloud-based data handling aligns with ethical standards and legal frameworks [1].

3.2 Banking Sectors

The banking sector is highly dependent on cloud computing for data storage, transaction processing, and customer services. However, financial institutions face significant cybersecurity threats, including data breaches, ransomware attacks, and regulatory compliance challenges. A study by [24] highlights the increasing risks associated with digital banking adoption, emphasizing the need for robust cybersecurity frameworks. Financial institutions must comply with stringent regulations such as the GDPR in Europe and the Cybersecurity Law in China. These regulations mandate strict data protection measures, requiring banks to ensure that customer financial data remains secure and accessible only within authorized jurisdictions. Additionally, the rise of AI and machine learning in financial services has introduced new vulnerabilities, necessitating continuous monitoring and risk assessment [25].

Geopolitical risks also play a crucial role in cloud adoption within the banking sector. The reliance on foreign cloud providers raises concerns about data sovereignty and potential government access to sensitive financial information. Countries are increasingly exploring sovereign cloud solutions to mitigate these risks and maintain control over their financial data [26].

3.3 Telecom Sector

The telecommunications industry is undergoing a rapid transformation with the deployment of 5G networks and cloud-based infrastructure. While these advancements enhance connectivity and service delivery, they also introduce data privacy concerns and national security risks. Governments worldwide have imposed restrictions on foreign telecom providers to safeguard critical communication networks [27].

Cloud computing enables telecom companies to manage vast amounts of data efficiently, but it also necessitates stringent security measures to prevent unauthorized access and cyber espionage. Cloud-based 5G networks raise concerns about foreign access to sensitive data, prompting regulators to enforce stricter data protection laws. Furthermore, telecom providers must navigate cross-border regulations that impact data storage and transmission. The European Union's GAIA-X initiative aims to establish a sovereign cloud infrastructure that ensures data privacy and security for telecom operators. Similar efforts are being made in other regions to reduce dependency on foreign cloud providers and enhance national cybersecurity resilience [28].

3.4 Healthcare Sector

The healthcare sector is increasingly adopting cloud computing to enhance efficiency in electronic health records (EHRs), medical research, and telemedicine. Cloud-based solutions offer benefits such as cost savings, scalability, and improved data accessibility. However, concerns surrounding cybersecurity, geopolitical risks, and ethical considerations remain critical for healthcare providers and policymakers. For instance, Healthcare institutions store sensitive patient data in cloud environments, making them prime targets for cyberattacks [29]. [30] highlights the rising frequency of cyber threats targeting healthcare infrastructure, emphasizing the need for stronger security measures. Unauthorized data breaches can compromise patient privacy, disrupt medical services, and pose significant legal consequences.

As healthcare systems increasingly rely on cloud computing, cross-border data sharing introduces geopolitical risks. Different countries have varying regulations regarding health data protection, complicating international collaboration. The European Union enforces the GDPR, ensuring stringent controls on patient data transfers, while other nations have differing policies that may conflict with global healthcare initiatives. Additionally, concerns about foreign cloud providers storing medical AI datasets raise questions about national data sovereignty and security. In fact, ethical concerns surrounding cloud computing in healthcare include patient consent, equitable access, and data security [31]. Ensuring informed consent for patients whose data is stored and analyzed on cloud platforms is essential to maintaining trust. As cloud-based services become standard, disparities in healthcare access may widen, especially for populations with limited technological infrastructure. Additionally, securing cloud-based healthcare records requires robust encryption and authentication mechanisms to prevent unauthorized access [32].

3.5 Military

Cloud computing is revolutionizing military operations by enabling secure data storage, intelligence analysis, and battlefield communication. Military organizations increasingly rely on cloud-based solutions to enhance operational efficiency, decision-making, and cybersecurity resilience. However, national security risks, cyber warfare, and geopolitical tensions make military cloud adoption highly sensitive. The integration of cloud computing in military operations introduces cybersecurity vulnerabilities that adversaries may exploit. Cyber warfare has emerged as a critical threat to national security, encompassing attacks on military networks, intelligence systems, and defense infrastructure. Such attacks can compromise classified data, disrupt operations, and weaken national defense capabilities [33].

The U.S. Department of Defense's Joint Enterprise Defense Infrastructure (JEDI) cloud project faced scrutiny due to concerns about data security and foreign interference. The project aimed to modernize military cloud infrastructure but raised questions about data sovereignty and the risks associated with relying on commercial cloud providers. Governments are now prioritizing sovereign cloud solutions to mitigate these risks and ensure secure military data management. Geopolitical tensions influence military cloud adoption, particularly in the context of cyber espionage and cloud-based warfare strategies. Adversaries may leverage cloud vulnerabilities to conduct cyber espionage, steal

sensitive defense data, and disrupt military communications. The increasing reliance on AI-driven cloud systems for intelligence analysis and autonomous military operations further complicates security concerns [34].

Cloud computing also supports digital warfare preparedness, helping nations develop cyber deterrence strategies against state-sponsored intrusions. AI-enhanced threat intelligence platforms enable real-time cyber threat analysis, adversarial behavior prediction, and automated incident response [35]. However, challenges such as data sovereignty enforcement and cloud infrastructure security must be addressed to maximize the effectiveness of cloud-based cyber defense strategies. The use of AI-driven cloud systems in military applications raises ethical concerns related to autonomous weapons, surveillance ethics, and military transparency. Autonomous weapons powered by cloud-based AI algorithms pose moral dilemmas regarding accountability and decision-making in combat scenarios. Ensuring ethical oversight in AI-driven military operations is essential to prevent unintended consequences and uphold international humanitarian laws. Surveillance ethics also arise, as cloud computing enables vast data collection and intelligence gathering, requiring a careful balance between national security and individual privacy. Additionally, military transparency is crucial in maintaining public trust and ensuring responsible cloud adoption in defense strategies [33].

4 Conclusion

While cloud computing offers numerous advantages across various sectors—such as government services, banking, healthcare, telecommunications, and military operations—it also presents significant geopolitical and ethical challenges that must be addressed. This aligns with [5] and is further supported by findings from [36], who demonstrated how cloud infrastructures have become strategic assets in global data governance competitions.

Data sovereignty concerns have prompted nations to develop sovereign cloud infrastructures to protect sensitive information, emphasizing the importance of secure, transparent, and resilient systems. [5] defined data sovereignty as the control and geographical ownership of data, underscoring its role in protecting sensitive information. In this context, [36] argued that technical measures like encryption and localized data centers are critical to achieving sovereignty.

Ethical considerations, including privacy, transparency, and accountability, are crucial in ensuring responsible cloud adoption that upholds human rights and fosters public trust. [37] claimed that cloud providers must address privacy, transparency, and accountability, not only as ethical imperatives but to align with frameworks like the UN Guiding Principles on Business and Human Rights. Similarly, [38] emphasized ethical considerations as essential to fostering trust and responsible adoption, with a focus on privacy compliance and data access rights.

As governments and industries continue to leverage cloud technology for innovation and efficiency, balancing these benefits with risks related to cybersecurity, geopolitical tensions, and ethical responsibilities will be essential. [39] asserted that public sector organizations, while increasingly adopting cloud technologies, they face complex challenges that extend beyond technical concerns. Key issues such as cybersecurity vulnerabilities, data governance issues, and geopolitical risks tied to cloud service providers

highlight the need for a balanced approach. Ensuring secure and sovereign cloud adoption requires robust governance frameworks that address operational efficiency, national security and ethical accountability.

Global strategies towards cloud computing reflect diverse perspectives and priorities. The U.S. plays a leading role in shaping international standards, yet faces complex challenges related to data sovereignty and ethical governance. Europe's emphasis on privacy and data autonomy, exemplified by regulations like GDPR and initiatives such as GAIA-X, underscores its commitment to digital sovereignty and high ethical standards. Meanwhile, regions like the Middle East seek to balance digital sovereignty with economic diversification and regional cooperation, navigating geopolitical intricacies. China's security-first approach highlights the importance of data sovereignty and state control, though it also raises ethical concerns around privacy and corporate responsibility. These varied approaches demonstrate the necessity for international cooperation and robust policy frameworks that align innovation with security and ethics. This aligns with the perspective presented by [40], which emphasizes the growing importance of international cooperation, with both policymakers and multinational enterprises playing key roles.

The future of cloud computing could depend on establishing resilient, transparent, and ethically grounded infrastructures worldwide. Policymakers, industry leaders, and ethical scholars must collaborate to address emerging challenges, from cybersecurity risks to safeguarding individual rights, fostering responsible digital transformation. Developing comprehensive international regulatory frameworks and standards could be vital in managing geopolitical tensions and ethical dilemmas. As technology advances, a balanced approach—prioritizing security, privacy, and ethical responsibility—could be critical to ensuring that cloud computing continues to benefit society while minimizing risks and fostering global trust in this transformative technology.

References

1. Tang, M.: The challenge of the cloud: between transnational capitalism and data sovereignty. In: *The Geopolitics of Chinese Internets*, pp. 63–77. Routledge (2024)
2. Murphy, B., Rocchi, M.: Ethics and cloud computing. In: Lynn, T., Mooney, J.G., van der Werff, L., Fox, G. (eds.) *Data Privacy and Trust in Cloud Computing*, 6. Palgrave Macmillan (2021). https://doi.org/10.1007/978-3-030-54660-1_6
3. Faragardi, H.R.: Ethical considerations in cloud computing systems. *Proceedings* 1(3), 166 (2017). <https://doi.org/10.3390/IS4SI-2017-04016>
4. Folorunso, A., Adewa, A., Babalola, O., Nwatu, C.E.: A governance framework model for cloud computing: role of AI, security, compliance, and management. *WJARR* 24(2), (2024). <https://doi.org/10.30574/wjarr.2024.24.2.3513>
5. Hummel, P., Braun, M., Tretter, M., Dabrock, P.: Data sovereignty: a review. *Big Data & Society* 8(1) (2021). <https://doi.org/10.1177/2053951720982012>
6. Lehdonvirta, V., Wú, B., Hawkins, Z.: Weaponised interdependence in a bipolar world: how economic forces and security interests shape the global reach of US and Chinese cloud data centers. *Rev. Int. Polit. Econ.* 1–26 (2025). <https://doi.org/10.1080/09692290.2025.2489077>
7. Raja, V., Chopra, B.: Exploring challenges and solutions in cloud computing: a review of data security and privacy concerns. *J. Artif. Intel. Gene. Sci.* 4(1), 121–144 (2024). <https://doi.org/10.60087/jaigs.v4i1.86>

8. Akpuokwe, C.U., Bakare, S.S., Eneh, N.E., Adeniyi, A.O.: Corporate law in the era of globalization: a review of ethical implications and global impacts. *J. Fin. Acc. Res.* **6**(3), 304–319 (2024). <https://doi.org/10.51594/farj.v6i3.857>
9. Vassilakopoulou, P., Hustad, E.: Bridging digital divides: a literature review and research agenda for information systems research. *Inf. Syst. Front.* **25**, 955–969 (2023). <https://doi.org/10.1007/s10796-020-10096-3>
10. Raihan, M.M.H., Subroto, S., Chowdhury, N., Koch, K., Ruttan, E., Turin, T.C.: Dimensions and barriers for digital (in)equity and digital divide: a systematic integrative review. *Digi. Trans. Soc.* **4**(2), 111–127 (2025). <https://doi.org/10.1108/DTS-04-2024-0054>
11. Ünver, H.A.: Politics of digital surveillance, national security and privacy. Centre for Economics and Foreign Policy Studies (2018). <http://www.jstor.org/stable/resrep17009>
12. European Data Protection Board, https://www.edpb.europa.eu/news/news/2023/edpb-determines-privacy-recommendations-use-cloud-services-public-sector-adopts_en. Last accessed 07 June 2025
13. Autolitano, S., Pawlowska, A.: Europe's quest for digital sovereignty: GAIA-X as a case study. Istituto Affari Internazionali (IAI) (2021). <https://www.jstor.org/stable/resrep30940>
14. European Parliament. https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI%282019%29640163_EN.pdf. Last accessed 07 June 2025
15. Mosa, M.J., Barhoom, A.M., Alhabbash, M.I., Harara, F.E.S., Abu-Nasser, B.S., Abu-Naser, S.S.: AI and ethics in surveillance: balancing security and privacy in a digital world. *Int. J. Acad. Eng. Res.* **8**(10), 8–15 (2024)
16. Connolly, L.Y., Lang, M., Giboney, J.: A study of the privacy paradox amongst young adults in the united arab emirates. SSRN. (2025). <https://doi.org/10.2139/ssrn.5166978>
17. Alazab, M., Alazab, M., Shalaginov, A., Mesleh, A., Awajan, A.: Intelligent mobile malware detection using permission requests and API calls. *Futur. Gener. Comput. Syst.* **107**, 509–521 (2020). <https://doi.org/10.1016/j.future.2020.02.002>
18. Albreem, M.A., Sheikh, A.M., Bashir, M.J., El-Saleh, A.A.: Towards green Internet of Things (IoT) for a sustainable future in Gulf Cooperation Council countries: Current practices, challenges and future prospective. *Wireless Netw.* **29**(2), 539–567 (2023). <https://doi.org/10.1007/s11276-022-03133-3>
19. Liu, J.: China's data localization. *Chinese J. Comm.* **13**(1), 84–103 (2019). <https://doi.org/10.1080/17544750.2019.1649289>
20. Creemers, R.: Cybersecurity law and regulation in China: securing the smart state. *China Law and Society Review* **6**(2), 111–145 (2023). <https://doi.org/10.1163/25427466-06020001>
21. Ong, R.: Privacy and personal information protection in China's all-seeing state. *Int. J. Law and Info. Technol.* **31**(4), 349–375 (2023). <https://doi.org/10.1093/ijlit/eaac003>
22. Gu, H.: Data, Big Tech, and the new concept of sovereignty. *J. Chin. Polit. Sci.* **29**, 591–612 (2024). <https://doi.org/10.1007/s11366-023-09855-1>
23. Mitchell, A.D., Samlidis, T.: Cloud services and government digital sovereignty in Australia and beyond. *Int. J. Law and Info. Technol.* **29**(4), 364–394 (2021). <https://doi.org/10.1093/ijlit/eaac003>
24. Cele, N.N., Kwenda, S.: Do cybersecurity threats and risks have an impact on the adoption of digital banking? *J. Fin. Crime* **32**(1), 31–48 (2025). <https://doi.org/10.1108/JFC-10-2023-0263>
25. Jain, R., Prajapati, D., Dangi, A.: Transforming the financial sector: a review of recent advancements in FinTech. *SSRN Electronic Journal* **8**(2) (2023). <https://ssrn.com/abstract=4380348>
26. Scott, H.S., Gulliver, J., Nadler, H.: Cloud computing in the financial sector: a global perspective. *Program on Int. Fin. Sys.* (2019). <https://ssrn.com/abstract=3427220>

27. Jafor, A.H.M., et al.: The evolution of cloud computing & 5G infrastructure and its economical impact in the global telecommunication industry. *AIJMR* **2**(5) (2024). <https://doi.org/10.62127/aijmr.2024.v02i05.1100>
28. Okori, W., Buteraba, S.: Cyber security exploits and management in telecommunication companies: the case of Uganda. *J. Comp. Sci. Technol. Stud.* **6**(4), 67–76 (2024)
29. Reegu, F.A., Abas, H., Gulzar, Y., Xin, Q., Alwan, A.A., Jabbari, A., et al.: Blockchain-based framework for interoperable electronic health records. *Sustainability* **15**(8), 6337 (2023). <https://doi.org/10.3390/su15086337>
30. World Economic Forum: <https://www.weforum.org/publications/global-risks-report-2023/>. Last accessed 07 June 2025
31. Corrales Compagnucci, M., Fenwick, M.: A multidisciplinary perspective on cross-border health data transfers: privacy, risks and solutions. In: *International Transfers of Health Data: A Global Perspective*, pp. 1–15 (2025). https://doi.org/10.1007/978-981-97-9983-1_1
32. Gadde, S., Amutharaj, J., Usha, S.: A security model to protect the isolation of medical data in the cloud using hybrid cryptography. *J. Info. Sec. Appl.* **73**, 103412 (2023). <https://doi.org/10.1016/j.jisa.2022.103412>
33. Sahu, K., Kumar, R., Srivastava, R.K., Singh, A.K.: Military computing security: insights and implications. *J. Inst. Eng. (India): Series B* (2024). <https://doi.org/10.1007/s40031-024-01136-6>
34. Digmelashvili, T.: The impact of cyberwarfare on national security. *Future Human Image* **19**(1), 12–19 (2023). <https://doi.org/10.29202/fhi/19/2>
35. Ofili, B.T., Ezeadi, S.C., Jegede, T.B.: Securing U.S. national interests with cloud innovation: data sovereignty, threat intelligence, and digital warfare preparedness. *Int. J. Sci. Res. Archive* **12**(1), 3160–3179 (2024). <https://doi.org/10.30574/ijrsra.2024.12.1.1158>
36. Timmers, P.: Sovereignty in the Digital Age. In: Werthner, H., et al. (ed.) *Introduction to Digital Humanism*, pp. 571–592. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-45304-5_36
37. Krishnamurthy, V.: With Great (Computing) Power Comes Great (Human Rights) Responsibility: Cloud Computing and Human Rights. *Bus. Human Rights J.* **7**(2), 226–248 (2022). <https://doi.org/10.1017/bhj.2022.8>
38. Yang, K., Yu, J., Lu, Y., Yang, L.T.: Security and privacy in cloud computing: technical review. *Future Internet* **14**(1), 11 (2022). <https://doi.org/10.3390/fi14010011>
39. Ukeje, N., Gutierrez, J., Petrova, K.: Information security and privacy challenges of cloud computing for government adoption: a systematic review. *Int. J. Inf. Secur.* **23**(2), 1459–1475 (2024). <https://doi.org/10.1007/s10207-023-00797-6>
40. Trasberg, K., Lundgren, M.: Unravelling cross-country regulatory intricacies of data governance: the relevance of legal insights for digitalization and international business. *J. Int. Bus. Policy* **7**, 112–127 (2023). <https://doi.org/10.1057/s42214-023-00172-1>

Society: Social Networks, Youth, and Gender in our Digital Society



The Factor of Social Media and Artificial Intelligence in the Radicalization of Young People in Jihadist Terrorism

Paula M. Núñez-Guerra^(✉) 

Complutense University of Madrid, Madrid, Spain
paulamnu@ucm.es

Abstract. This article is presented with the main objective of this research being to examine how terrorist organisations have adapted to new digital platforms and use major social media networks and artificial intelligence (AI) to reach an increasingly younger audience. The methodological approach adopted is descriptive in nature, based on an exhaustive bibliographic analysis of academic sources and research studies. It explores digital propaganda strategies, the use of recommendation algorithms, and segmentation techniques that enable the effective dissemination of extremist content. The scope of the study covers global platforms, focusing on the mechanisms that facilitate continuous exposure to radical messages, starting with the main question of “How do terrorist organisations penetrate the minds of young people who are recruited and radicalised?”.

Keywords: AI · social media · jihadist terrorism · radicalization

1 Introduction

One of the main challenges and threats to international security lies in jihadist terrorism. It is a threat that is becoming increasingly latent and is increasingly expanding to the international scene. As for Europe, the continent has been the scene of major jihadist attacks carried out by the two major terrorist organisations that ravage the international scene: Al Qaeda and Daesh. This was the case of 11 M in Madrid or the attacks that took place in Paris in 2015 and the fatal attack on Las Ramblas in Barcelona two years later. However, it is not the only cause of terrorism at an international level. Currently, the panorama is very diverse, but with a common denominator and that is to put the spotlight on the use of Artificial Intelligence and social networks by terrorist groups to recruit and radicalise new combatants. In five points, Pisabarro explains how the TikTok platform is a strategy to expand cyberjihad among young people: (1) due to its format based on creativity and “hooking” users through challenges, (2) due to its diversity in reaching any community, (3) due to its influence on public opinion, (4) due to young people’s fear of missing out on what happens on social media, and (5) due to the Artificial Intelligence algorithm that makes the user consume similar content.

According to Rapoport, from the end of the 19th century to the present, there are four waves of terrorism that the author differentiates. The first is led by anarchist terrorism

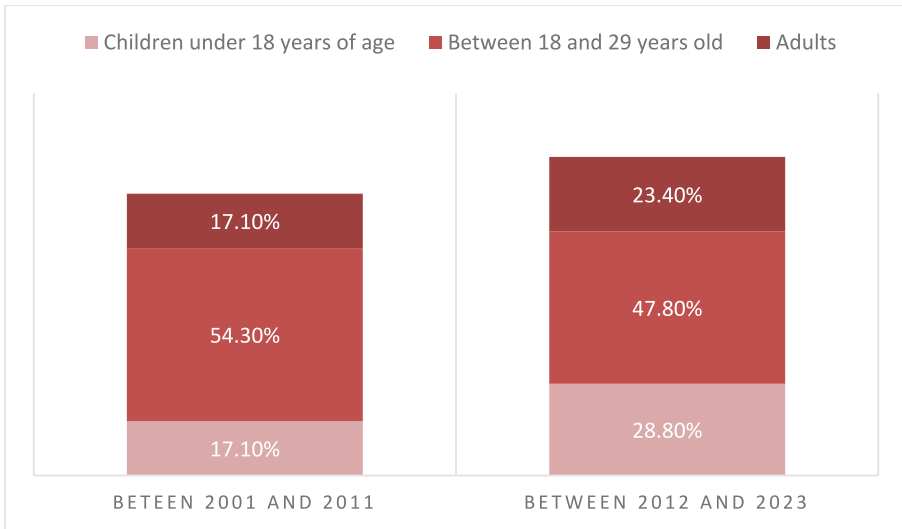
where the target of the attacks were the heads of state because they were seen as the cause of all the evils in society. This first typology continued until the time of Versailles, when, after this, the author indicates that the second wave began: “the colonial wave” where the objective of the attack was directed towards the security forces and bodies because, as the author explains, they were perceived as the ears and eyes of the State. From then until the 70s when the third wave described by the author begins: “the wave of the extreme left” where violence has no political weight and ends in the 80s with the fourth wave understood as “religious terrorism”. It is in this last one that we are currently in and that, according to the author, radicalization began to mutate into terrorism. These ideas together with the international impact caused by the 9/11 attack, demonstrated that far from this strictly local conception of terrorism, terrorism is an issue that concerns the international community.

Similarly, in reference to what Neumann explains, there are two types of terrorism: an “old terrorism” based on a counterattack structure where objectives that are understood as legitimate are pursued; and a “new terrorism” made up of a network structure whose purpose is to reach the international level and provoke mass attacks with excessive violence. However, within jihadist terrorism, authors such as Reinales, García and Vicente divide this phenomenon into three periods: a first that runs until months after 9/11, a second that spans a decade until 2011, coinciding with the bite of the then leader of Al Qaeda, Osama Bin Laden; and a third moment that begins with the jihadist insurgency in the context of the Syrian Civil War and which, in turn, coincides with the terrorist explosion on social networks. Therefore, not only was there talk of terrorism, but there was already talk of cyberterrorism.

The Internet itself provides a host of benefits and has changed the way we see the world, creating this information dependency. With a click of a button, we can find out what is happening in another part of the planet, participate virtually in events or get in touch with people miles away. But at the same time, the arrival of the Internet makes it possible for illegal actors to see this as an opportunity to commit their crimes, recruit and radicalize new combatants. Among them, it is worth highlighting the possibility of reaching new followers and radicalizing lone wolves. The current threat from the European Union in this matter is such that, in fact, according to Espinosa, most of the lone wolves who attack on European soil are citizens of the region.

This increasingly perennial idea of cyberjihad, together with what Vicente has stated, indicates that 23.40% of the jihadists convicted or killed in Spain at the beginning of the radicalisation process between 2012 and 2023 were under 18 years of age, which represents an increase of 6.3% points compared to the period between 2001 and 2011 (Fig. 1).

This is why this research will analyse how terrorist organisations target young people through social media, what messages they send to their target audience and through what formats. In this way, it will be observed how the role of young people in terrorist organisations has evolved and how they have used social media to expand their narrative to generation Z. Therefore, the main objective of this research is to analyse how terrorist organisations have adapted to the new platforms and are using the main social networks to reach an increasingly younger audience. This will either refute or confirm the initial hypothesis: terrorist organisations have adapted their propaganda machinery with



Own elaboration based on Vicente

Fig. 1. Jihadists convicted or killed in Spain, according to the period of detention or death and age group at the start of the radicalisation process.

social networks and Artificial Intelligence to attract and radicalise new and increasingly younger combatants. In addition, this research will look at the factor of the ethics of terrorism, taking into consideration authors such as Malatji and Tolah.

Therefore, under this idea and with the premise that more and more young people and minors are becoming radicalised through social networks, the main question will be answered: How do terrorist organisations penetrate the minds of young people who are recruited and radicalised?; which in turn will answer another question: What messages and formats do terrorist organisations use to attract new followers? All of this from the use of a methodology based on a descriptive method, thus allowing us to show the theoretical and conceptual framework about this phenomenon. After this, an analytical method has been carried out, which has allowed us to understand and establish the causal relationships between the transformations of jihadist propaganda and the growth of increasingly younger users on the main social networks. For both methods, consideration has been given to the views of Caldach, who indicates that the main objective of the descriptive method is to present the analysis under study in the most rigorous manner possible. To locate the documents analysed, I consulted academic repositories and searched for documents using keywords such as “social networks”, “artificial intelligence” and “terrorism”. Once I obtained the search results, I focused on two issues: publication dates and summaries. Apart from those documents that provide a theoretical basis, the articles used date from 2015 onwards, coinciding with the self-proclaimed caliphate by Daesh and, therefore, a greater “propaganda machine”. After reading the abstracts of each article, I selected those that would help guide the writing of the article and complement each other.

Furthermore, according to the Caldach [25], the analytical method “starts from a general understanding of a reality in order to distinguish, understand and classify the different essential elements that form part of it and the relationships between them”¹.

2 Social Networks and Jihadist Terrorism

As cyber media have evolved, the tools used by terrorist organisations to spread the idea of “war jihad” have changed significantly. Montes notes that the first tools were simple websites, which later evolved into forums and social networks such as Facebook. To avoid detection, terrorist organisations began to use more encrypted DeepWeb tools, such as ZeroNet. In addition, Zelin distinguishes four phases in the evolution of jihadist activity on the Internet. The first phase, in 1984, focused on ideological propaganda through sermons, leaflets, magazines, videotaped lectures and newsletters. In the second phase, in the mid-1990s, organisations created their own websites, where they decided what to publish and disseminate to attract international attention. The third phase, in the mid-2000s, saw the emergence of interactive forums, where Internet users also contributed content. The fourth phase, in the late 2000s, was characterised using social networks such as Facebook, Twitter and YouTube.

The propaganda campaigns of terrorist organisations, according to Expósito, have three main objectives: 1) Recruitment: targeting vulnerable groups, especially minors, using formats such as cartoons, music videos and games to capture their attention; 2) Initiation: implicitly encouraging the commission of terrorist acts; 3) Radicalisation: they indoctrinate recruits, transforming them into mujahideen committed to the cause. This is why Morán stresses that ICTs have become key tools for propaganda, internal communications, training, indoctrination, financing, radicalisation, recruitment and information gathering by terrorist groups. Jihadist terrorism has exploited propaganda as a key factor in its actions. According to an analysis mentioned by Sánchez, more than half of those surveyed attribute the “terror” factor to cyberterrorism. Morán also points out that the use of ICTs for terrorist and criminal purposes has spread rapidly since the beginning of the 21st century, explaining the expansion and emergence of new criminal groups. According to Lejarza, the success of the use of social networks among terrorist groups is due to several factors: 1) Publishing is cheaper than in conventional media; 2) Easy and non-centralised access, as only one device is needed; 3) Messages free from the controls and policies of traditional media; 4) Greater geographical dissemination than the press, radio or television; and 5) Immediacy in the publication of events. All these factors, as the author explains, make social networks effective platforms for distributing ideological propaganda in various formats such as magazines, forums, audio and video messages, offering ideological instruction, explanations, justifications and promoting terrorist activities (UN, cited in Lejarza).

In all these formats, López, Mendieta and Micó identify the main themes of jihadist propaganda: 1) criticism of the West, 2) defence and apology of jihad, and 3) manipulation of religious identity. This content, which includes extreme violence and weapons, is

¹ Translation by the autor from the original: “parte del conocimiento general de una realidad para realizar la distinción, conocimiento y clasificación de los distintos elementos esenciales que forman parte de ella y de las relaciones que mantienen entre sí”.

distributed through social networks. To this end, Bourekba explains that terrorist organisations use images and videos to give greater visibility to their content, fulfilling two functions: spreading terror and sowing confusion in public opinion, and deploying propaganda that informs about the military and state advances of the IEO, as well as seducing and recruiting new people. Social media has become a “virtual training ground” that allows groups to reach out to their audience, connect with young audiences and keep some of their heroes alive (Fanjul).

Having explained the above, it is worth noting, according to Expósito, that the main motivations for joining jihadism include the need to belong, identity crisis, social exclusion, attraction to the values of the terrorist group, the search for prestige and the need for communication. Thus, according to the author, young people of Muslim immigrant descent in the West, with no emotional ties to their country, are especially vulnerable. In terms of recruitment, if the targets are minors, as Lejarza [6] explains, terrorist organisations “use cartoons, music videos or computer games, offered in multiple languages, which include messages that promote and glorify terrorism”². Regarding women and girls, the author explains that some Western teenage girls aspiring to be prominent figures in Daesh’s jihadist propaganda have been targeted.

The profiles of individuals recruited by terrorist organisations are diverse and vary according to the context and recruitment strategies used. According to González, there is no such thing as a “typical terrorist”, but rather there are several profiles that share some common traits. Among the profiles recruited by terrorist organisations, the author distinguishes:

1. Strike teams: groups or cells arriving from abroad to carry out terrorist actions. An example is the pilots and hijackers responsible for the 9/11 attacks.
2. Born in Islamic countries and radicalised in the West: individuals of Muslim faith who have become radicalised after residing in Western countries. An example is the Hamburg cell, led by Mohammed Atta, composed of Muslim students who were radicalised in Germany.
3. Western-born and radicalised Muslims: young people born and socialised in Western countries who do not feel an emotional attachment to their country of birth. These individuals, often descendants of Muslim immigrants, find the concept of a nation of Islam proposed by terrorist groups such as Al Qaeda or Daesh appealing.
4. Integris converts: individuals who convert to Islam and adopt a radical ideology. These converts often seek a sense of belonging and purpose that they find in extremist ideology.
5. Recruiters: individuals with combat experience and a strong religious ethos, selected for their charisma and leadership skills. These recruiters are essential for influencing and recruiting other individuals.

The use of the internet and social media has facilitated this process, allowing recruiters to reach a wider and more diverse audience. Daesh, for example, has developed radicalisation and recruitment processes that include several methods like cyberspace.

² Translation by the autor from the original: “utilizan dibujos animados, videos musicales o juegos de ordenador, ofrecidos en múltiples idiomas, que incluyen mensajes que promueven y glorifican el terrorismo”.

These methods allow terrorist organisations to adapt their strategies to different contexts and profiles, increasing their capacity for recruitment and radicalisation. Social media such as Facebook, Twitter and YouTube have become key tools for the propaganda and mobilisation of young terrorists. These platforms allow recruiters to reach a wide and diverse audience, taking advantage of ease of access and the lack of effective controls.

Social media allows terrorist organisation to tailor their messages to different audiences and contexts. They use a variety of formats, such as videos, images, forums and audio messages, to engage and radicalise young people. This content is designed to be accessible and engaging, using visual and narrative elements that resonate with young people's interest and concerns.

In short, the internet and social media play a key role in the recruitment and radicalisation of individuals by terrorist organisations. These platforms allow for the rapid and massive dissemination of propaganda, facilitating the recruitment of new members and the spread of extremist ideology. The adaptability and flexibility of recruitment methods in cyberspace make these tools particularly effective in attracting and radicalising young people around the world.

3 The New Era of Cyberterrorism: The Use of AI

This section will look at how generative artificial intelligence (GenAI) models can be exploited by terrorists. As these technologies advance, so do concerns about their misuse to enhance terrorist operations both online and in the real world. Considering Weimann, Pack, Sulciner, Sheinin, Rapoport and Diaz, the main risk of Generative AI in this context could be grouped under three headings:

1. **Propaganda and Recruitment.** According to the authors, generative AI models can create highly realistic multimedia content, including text, images, videos and audio. This allows terrorists, as the authors point out, to produce convincing and personalised propaganda, facilitating the recruitment and radicalisation of individuals. Furthermore, according to Ruiz (2024), terrorist groups use AI to automate the creation of propaganda content. This includes, according to the author, the generation of avatar videos and the automatic translation of messages into multiple languages, which allows them to overcome language barriers and increase the distribution of their messages globally. In addition, the author indicates that chatbots with AI capabilities are used to provide personalised information to potential recruits, tailoring messages according to their beliefs and interests.
2. **Planning and Execution of Operations.** According to the authors, generative AI can be used to plan and execute terrorist operations with greater precision and efficiency. This ties in with the authors' idea that deep learning algorithms can analyse large volumes of data to identify vulnerabilities and optimise attack strategies. This includes, as the authors point out, the automation of complex tasks such as target identification and target coordination and the coordination of cyber-attacks. In this section, Ruiz states that machine learning algorithms can analyse large volumes of data to identify vulnerabilities in computer systems and plan cyber attacks with greater accuracy. This, according to the author, allows terrorists to operate at unprecedented scale and speed. Indeed, the author notes that one of the most worrying applications of AI is the use of

autonomous drones. Such drones, the author describes, can be programmed to carry out reconnaissance missions and attacks without human intervention. The integration of AI into drones, as the author explains, improves autonomous navigation, target recognition and real-time mission planning, increasing the lethality of attacks.

3. Deepfakes and Disinformation. According to the authors, deepfakes, created through generative AI, can be used to spread disinformation and sow chaos. Such actions aim, as Ruiz argues, to sow chaos and confusion, as well as manipulate public opinion to amplify their message.

Having explained this, more specifically, Schaer states that a group affiliated with Al Qaeda announced online workshops on AI and published a guide on the use of AI chatbots, which shows an effort on the part of the organisation to educate its followers in the use of these technologies. On the other hand, and under the umbrella of propaganda and disinformation, the author exposes that after the attack in Moscow, a follower of Daesh created and spread a fake news story about the attack. In this sense and as indicated, AI is used to generate content that can go unnoticed by social media moderators. In particular, Daesh uses generative AI tools to create convincing multimedia content, such as deepfakes, videos and audios. This content is designed to appear authentic and credible, which, according to the author, increases its ability to influence target audiences. An example of this was a manipulated video from the series “Family Guy” in which the main character sings a song with extremist lyrics: “Our weapons are heavy, our ranks are many, but Allah’s soldiers are more than ready” (Schaer).

In addition, according to the author, AI enables the automation of propaganda creation and distribution. In this sense, the author indicates that chatbots and other automated systems can interact with potential recruits, personalising messages according to their interests and beliefs. This, as the author points out, facilitates radicalisation and recruitment without the need for direct human intervention. This use of AI complements Al Qaeda and Daesh propaganda, making messages harder to detect.

In short, terrorist organisations use artificial intelligence (AI) to target various specific groups to maximise their impact and recruitment. Taking the above into account, one could distinguish six main groups of audiences that terrorist organisations target with the use of AI:

1. Disaffected and Vulnerable Youth. Young people who feel marginalised, alienated or dissatisfied with their current circumstances are a prime target. These individuals may be searching for purpose or identity, and extremist ideologies can offer them a sense of belonging and meaning.
2. People in Economic and Social Hardship. Those facing economic, social or political hardship are more susceptible to radicalisation. Terrorist organisations exploit these vulnerabilities to attract individuals looking for a way out of their problems.
3. Social Media Users. Social media is a powerful tool for terrorist organisations. They use AI algorithms to identify and segment users who are most susceptible to radicalisation and then bombard them with extremist content. AI chatbots can interact with these users, tailoring messages according to their personal interests and concerns.
4. Marginalised communities. Communities that feel marginalised or discriminated against may be more receptive to messages from terrorist organisations that resonates with the experience and emotions of these communities.

5. Individuals with Specific Interests. Terrorists also use AI to target individuals with specific interests that can align with their ideologies. For example, as seen above, they can create content that appeals to people interested in political, religious or social justice issues, tailoring messages to be more persuasive.
6. People in Conflict Regions. People living in regions of conflict or political instability are especially vulnerable. Terrorist organisations use AI to disseminate propaganda that promises solutions to their problems and can appear more credible in contexts of desperation.

4 Conclusions

This research has provided an answer to the questions posed at the beginning and has managed to observe how terrorist organisations penetrate the minds of young people who are recruited and radicalised; in addition to being able to see what messages and formats terrorist organisations use to attract new followers. Furthermore, it has been possible to confirm the main hypothesis, which is that it has been possible to study how terrorist organisations have adapted their propaganda machinery to social networks and Artificial Intelligence. On the one hand, it has been possible to see that terrorist organisations have adapted their propaganda tools as cybernetic media have evolved, from multiple websites to social networks and encrypted DeepWeb tools. In this regard and as discussed above, Zelin identifies four phases in jihadist activity on the internet: ideological propaganda (1984), website creation (mid-1990s), interactive forums (mid-2000s). Propaganda campaigns, as we have seen, have three main objectives: recruitment, initiation and radicalisation. This is why ICTs are essential for propaganda, internal communications, training, indoctrination, financing, radio-qualification, recruitment and information gathering. In addition, social networks are effective because of their low cost, easy and decentralised access, uncontrolled messaging, wider geographical spread and immediacy. It has also been observed that the main themes of jihadist propaganda include criticism of the West, defence of jihad and manipulation of religious identity. Similarly, motivations for joining jihadism have been found to include the need to belong, identity crisis and social exclusion. Also, social networks allow messages to be tailored to different audiences, facilitating the recruitment and radicalisation of young people. Moreover, social networks are effective due to their low cost, easy and decentralised access, low cost of access, low cost of communication and low cost of access.

On the other hand, the findings on the use of Generative Artificial Intelligence (GenAI) models by terrorists show that several risks stand out. Instead, propaganda and recruitment are enhanced by GenAI's ability to create highly realistic multimedia content, facilitating the production of convincing and personalised propaganda. This includes the automatic generation of avatar videos and the translation of messages into multiple languages, overcoming language barriers and extending global distribution. Secondly, it has been observed that the planning and execution of operations are optimised using deep learning algorithms that analyse large volumes of data to identify vulnerabilities and coordinate attacks. In this sense, and as explained above, the integration of AI in autonomous drones improves navigation, target recognition and mission planning in real time, increasing the lethality of attacks. And finally, as we have seen,

deepfakes and disinformation represent a significant threat, as they enable the creation of false content that can sow chaos and manipulate public opinion. This is why, as we have seen, terrorists use these tools to generate content that goes unnoticed by social network moderators, thus increasing their capacity to influence.

In coalition with this, Malatji and Tolah examine the complex dimensions of artificial intelligence (AI)-driven cyberattacks. It presents the Arc of AI Cybersecurity Dimensions (AICD) designed to guide academics, policymakers and industry practitioners in understanding and combating the challenges posed by offensive AI. The study underscores the need for adaptive defences and considers ethical considerations crucial to mitigating the risks associated with adversarial AI. Through rigorous textual analysis and extensive literature reviews, the article highlights the urgency of interdisciplinary approaches to bridge the gap between technology and humanity in cybersecurity discussions.

Taking all the above into account, in short, this study is an example of an analysis of how terrorist organisations use social networks and artificial intelligence to recruit and radicalise young people, adapting their propaganda to advanced cyber media. It also highlights the use of deepfakes and disinformation to manipulate public opinion, and the optimisation of operation through deep learning algorithms and autonomous drones. In the same way, this research can open the way for future research on how to use this methodology for other cases or even open the possibility of future comparisons between different armed conflicts and the different obstacles to the judicialization of women in different armed conflicts.

References

1. Bourekba, M.: Por qué atrae el “Estado Islámico”. CIDOB (112), 1–5 (2015) <https://dialnet.unirioja.es/servlet/articulo?codigo=9207386>
2. Calduch, R.: Métodos y técnicas de investigación en Relaciones Internacionales. Universidad Complutense de Madrid, 1–161 (2020). <https://www.ucm.es/data/cont/media/www/pag-55163/2Metodos.pdf>
3. Espinosa, J.L.: Las peores masacres cometidas por los «lobos solitarios». ABC (2016). https://www.abc.es/internacional/abci-atentados-grande-grito-yihadista-precede-masacre-201605101506_noticia.html. Last accessed 19 December 2024
4. Expósito, J.: Reclutamiento yihadista en redes sociales: Yihad 3.0 Un análisis del impacto de las actividades de reclutamiento a través de las redes sociales. Cuadernos de la Guardia Civil: Revista de seguridad pública (62), 55–70 (2021). <https://dialnet.unirioja.es/servlet/articulo?codigo=7811970>
5. Fanjul, M.L.: El mensaje persuasivo radical: yihadismo y redes sociales. Instituto Español de Estudios Estratégicos (115), 1–14 (2015). <https://dialnet.unirioja.es/servlet/articulo?codigo=7687285>
6. González, M.I.: Perfil del terrorista y del reclutador yihadistas. Safe World (2022). <https://www.redsafeworld.net/post/perfiles-terroristas-reclutadores-yihadistas>
7. Lejarza, E.: Terrorismo islamista en las redes – la yihad electrónica. Instituto Español de Estudios Estratégicos (100), 1–18 (2015). <https://dialnet.unirioja.es/servlet/articulo?codigo=7687270>
8. López, S., Mendieta, G., Micó, J.S.: Tendencias online de la propaganda yihadista. Análisis del caso español. Instituto Español de Estudios Estratégicos (85), 1–21 (2021). <https://dialnet.unirioja.es/servlet/articulo?codigo=8153074>

9. Malatji, M., Tolah, A.: Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. *AI Ethics*, 1–28 (2024). <https://doi.org/10.1007/s43681-024-00427-4>
10. Montes, D.: A vueltas con el terrorismo e internet: hacia una definición de ciberterrorismo. *Revista de Derecho UNED* **27**, 697–738 (2021). <https://doi.org/10.5944/rduned.27.2021.31102>
11. Morán, S.: La ciberseguridad y el uso de las tecnologías de la información y la comunicación (TIC) por el terrorismo. *Revista Española de Derecho Internacional* **69**(2), 195–221 (2017). <https://doi.org/10.17103/redi.69.2.2017.1.08>
12. Neumann, P.: Old & New Terrorism. In: Neumann, P. (ed.) *Old & New Terrorism*, pp. 14–48. Polity Press, Cambridge (2009). https://www.politybooks.com/bookdetail?book_slug=old-and-new-terrorism--9780745643762
13. Pisabarro, S.: La radicalización yihadista de los jóvenes en Europa a través de TikTok. *Instituto Español de Estudios Estratégicos* (2), 1–24 (2024). <https://dialnet.unirioja.es/servlet/articulo?codigo=10120332>
14. Rapoport, D.C.: The Four Waves of Modern Terrorism. In: Cronin, A., Ludes, J. (eds.) *Attacking Terrorism*, pp. 46–73 (2005). <https://icct.nl/sites/default/files/import/publication/Rapoport-Four-Waves-of-Modern-Terrorism.pdf>
15. Reinares, F., García, C., Vicente, Á.: Yihadismo y yihadistas en España. Quince años después del 11M. *Real Instituto Elcano* (159), 1–200 (2019). <https://www.realinstitutoelcano.org/monografias/yihadismo-y-yihadistas-en-espana-quince-anos-despues-del-11-m/>
16. Ruíz, Ó.: Terrorismo e inteligencia artificial, un tándem mortal. *The Diplomat* (2024). <https://thediplomatinspain.com/2024/10/28/terrorismo-e-inteligencia-artificial-un-tandem-mortal/>
17. Sánchez, L.M.: Repensando el concepto de ciberterrorismo. *Instituto Español de Estudios Estratégicos* (11), 1–15 (2021). <https://dialnet.unirioja.es/servlet/articulo?codigo=7957072>
18. Schaer, C.: Estado Islámico usa inteligencia artificial para propaganda. *DW* (2024). <https://www.dw.com/es/estado-isl%C3%A1mico-y-el-peligroso-uso-de-inteligencia-artificial-para-su-propaganda/a-69638260>
19. Vicente, Á.: Radicalización yihadista en España: menores, espacios virtuales y la resonancia de conflictos internacionales. *Real Instituto Elcano* (31), 1–13 (2024). <https://www.realinstitutoelcano.org/analisis/radicalizacion-yihadista-en-espana-menores-espacios-virtuales-y-la-resonancia-de-conflictos-internacionales/>
20. Weimann, G., et al.: Generating terror: the risk of generative AI exploitation. *Combating Terrorism Center At West Point* (17), 1–41 (2024). <https://ctc.westpoint.edu/generating-terror-the-risks-of-generative-ai-exploitation/>
21. Zelin, A.: The State of Global Jihad Online. A Qualitative, Quantitative, and Cross-Lingual Analysis. *New America Foundation*, 1–24 (2013). https://static.newamerica.org/attachments/4360-the-state-of-global-jihad-online/Zelin_Global%20Jihad%20Online_NAF.453b75377e4f4588bac20fa7e57a9a8a.pdf



Using AI for Research and Educational Support Enhancing the Design of Computer-Based Evaluations

Joana M. Matos^{1,2,3,4}  and Adrian-Horia Dediu^{1,2,3} 

¹ Universidade Autónoma de Lisboa, Palácio Dos Condes Do Redondo, R. de Santa Marta 56,
1169-023 Lisboa, Portugal

{jmatos, adediu}@autonoma.pt

² Autónoma TechLab, 1169-023 Lisboa, Portugal

³ Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal

⁴ Nova School of Business and Economics, Rua da Holanda, no. 1, 2775-405 Carcavelos,
Portugal

Abstract. This paper has two objectives related to using AI generative tools for research and education: a) exploring the opinion of a group of students on this matter and b) sharing our experience in AI-assisted methods to design various question types for online evaluation. Teachers can ethically use AI, among other activities, to prepare courses and design enhanced online assessment question databases. In this way, we can generate a large enough database of questions to ensure a wide variety of topics are covered and prevent repetitions among different students' questions. We developed and implemented a Moodle quiz questionnaire for students and collected their opinions on using AI during their work. We analyzed the students' answers to understand their level of awareness regarding the ethical implications of using AI during academic activities. We share these results in the second part of our paper.

Keywords: Academic Integrity · Ethical AI use · Generation Z perspectives on AI

1 Introduction

Generative Artificial Intelligence has become important for knowledge creation, acquisition, and dissemination. However, using AI-generated text in academic writing raises ethical problems regarding authorship, introduces risks of plagiarism, limits the authors' creativity and capability to handle and structure ideas, and compromises research integrity [1].

Practically, there are no ethical concerns about using AI tools for literature search overviews and summaries. However, one should check the sources' reliability and relevance to the research subject. Sometimes, the results of this stage can also motivate the selection of a new topic for the project or research work.

There are a few situations where using AI-generated data for research purposes is acceptable, mostly when studying topics related to AI characteristics. Otherwise, the data source should be documented and acknowledged. Sometimes, researchers produce data via computer simulations and can benefit from AI's help during the modelling design and coding stages. The IEEE [2, 3] and ACM [4] codes provide foundational ethical guidance for integrity, honesty, respect for privacy, and professional responsibility that shape the processes by which data is acquired, managed, and reported.

Nevertheless, using acknowledged AI tools for grammar checking, proofreading, and style improvement is acceptable [5]. Generative AI may also help to analyze the clarity, structure, and understandability of academic writing.

Relying on Generative AI tools is one of the most relevant topics associated with studies about Generation Z, or Gen Z for short, which designates people born between the late 1990s and early 2010s. They are our youth today [6]. Many things were said to explain the characteristics of this generation. However, we would only refer to several known keywords that characterize this generation's everyday life, such as the Internet, diversity, individualism, pragmatic attitude, climate change, and valuing collaboration, which is just a glimpse of the complexity of such an approach.

This paper explores current ethical challenges when using Artificial Intelligence (AI)-based tools for the stages of academic writing. We collected the opinions of a group of bachelor students on these problems through a questionnaire. The students answered 10 quiz questions in Moodle. In the extended abstract, we briefly present and analyze the preliminary results of this research.

2 Related Work, Guidelines and Frameworks for AI Use in Higher Education

As generative AI technologies become a regular part of our daily activities, many governmental, intergovernmental, or independent organizations have issued various communications, ethical guidelines, and policy frameworks about AI applications. In particular, Higher Education Institutions (HEIs) play an active role in using and explaining the newly emerged paradigm. Opportunities and risks are associated with AI usage in learning, teaching, and assessment. We need institutional debates, followed by regimentations, to preserve academic integrity and align with broader societal values.

The Joint Information Systems Committee (JISC), a UK-based organization supporting digital solutions in higher education, has published several documents addressing generative AI usage in HEIs and its ethical implications. For example, its *AI maturity toolkit for tertiary education* helps universities assess their stage in AI adoption [7]. At the same time, their primer about *Generative AI* offers a concise overview of generative AI technologies and their implications for curriculum, assessment, and academic policies. The attribute associated with AI penetration in our activities is worth mentioning: “Generative AI is likely to be pervasive, so banning it is not an option” [8]. Their guide, *Embracing Generative AI*, advocates rethinking how exams are designed in an AI-aware context, encouraging institutions to move beyond proctoring and toward better question design to prevent misconduct by construction [9]. Prof. Phillip Grandinetti from Ohio State University has a similar perspective when discussing the 3Ds model:

Design, Deter, and Detect, which are the basic principles used when integrating AI in the student's assessment [10].

A large number of quiz questions, with new questions added before each test, to prevent repetitions among students or knowledge transmitted from previous years, sequential access to the test's questions, an appropriate amount of time per question, and some tolerance in case of mistakes are only a few of the basic principles we consider in the design stage of our tests. We can use AI to help the question design and to evaluate the difficulties of a test set, but we cannot trust AI to correct answers or to evaluate the use of AI by students. We can use AI to construct the feedback while evaluating the students' answers. An entire episode of the podcast "*Beyond the Technology*" dedicated to *Rethinking Assessment and Feedback* explains the new challenges in redesigning the current assessments. Simon Walker, Director of Program Development at University College London, invited by Sarah Knight, Head of Learning and Teaching Transformation at JISC, underlines the importance of understanding the fundamental aspects related to feedback, such as "feedback in terms of learning, feedback on assessment or the assessment that governs feedback" [11]. We agree with Simon, who says we do not understand these things well. However, we comment on an example of assessment that governs feedback. Deciding when to give the answers to students after a Moodle quiz depends on the weight of the mark in the general evaluation, the number of participants, and our logistic and test conditions. We can decide on immediate or delayed answers to allow some reflection time and reduce the possibility of sharing the answers with students still taking the test.

Our practical choices as educators are similar to broader debates on the responsible integration of technology in education. Documents from UNESCO (United Nations Educational, Scientific and Cultural Organization) share a common perspective on promoting AI in higher education institutions. Besides the introductory material [12], their 2023 book *Guidance for Generative AI in Education and Research* reveals several controversies, like using content without consent or the currently unexplainable models used by the generative tools. A significant cause comes from these models lacking the "understanding" of the real world. There are consequences: this technology facilitates the appearance of deepfakes and reduces diversity because the marginalized voices become even more insignificant. It discusses the requirements and possible risks of using generative AI as an advisor for outlines, literature review, teaching assistant, and coach for coding [13].

The OECD (Organization for Economic Cooperation and Development) also addresses the implications of AI in education, with a focus on equity, inclusion, and teacher empowerment. In their 2023 publication, *Opportunities, Guidelines and Guardrails for Effective and Equitable Use of AI in Education*, the OECD encourages HEIs to develop institutional policies that prioritize teacher control over AI use, address potential biases in datasets, and ensure that students from all backgrounds benefit equally from AI tools. The OECD further emphasizes the importance of training staff and students to use AI responsibly, suggesting that AI competence is becoming important in academic contexts [14]. This material is the result of the cooperation between OECD and Education International. They published a brochure introducing the discussions at the International Summit of the Teaching Profession in 2023 [15].

In contrast to the more pedagogically focused documents above, the European Commission's work—particularly the Ethics Guidelines for Trustworthy AI—takes a regulatory and legalistic attitude. It outlines the key requirements that AI systems should meet to be considered trustworthy. It presents a hierarchy of risks, classifying AI applications according to unacceptable, high-risk, limited, and minimal risk. We can find a high-level summary of the AI Act at [16], and the full text is available at [17].

Together, these publications represent the beginning of a broader response to the challenges and possibilities of introducing AI in higher education. Although many of their recommendations are still high-level or advisory, they mark an important step toward forming a shared vocabulary, setting minimum ethical standards, and encouraging proactive institutional planning. As we move forward, HEIs will adopt these guidelines and shape their internal policies, pedagogical practices, and assessment methods in alignment with these evolving frameworks.

We conclude this section by recalling Sue Attewell's JISC blog [18], which identifies three possible approaches to using generative AI in assessments: *avoid*, *outrun*, or *embrace*. These three approaches cover many possible responses, even though their boundaries are not crisp.

- **Avoid:** Notably, Attewell does not advocate a total rejection or absolute ban on AI.
- **Outrun** is not directly related to AI acceptance or rejection. It speculates on various known limitations of AI systems to discourage cheating. This strategy might yield results in the short term; however, new AI models could make our efforts in this direction worthless.
- **Embrace:** We acknowledge that the last strategy seems the most likely to follow. As Attewell notes, ensuring equitable access will be challenging.

3 Methodology

Following the principles of efficient online assessments discussed in the literature, particularly those recommending careful design, proactive deterrence, and detection, we explored AI generative tools to design diverse and complex Moodle quizzes [10].

Because our university struggles to buy advanced proctoring tools, we prioritized quiz design and deterrence measures over invasive detection technologies.

These questions' complexity and varied nature inherently enhance academic integrity by discouraging superficial or repetitive answers. Furthermore, the direct XML (eXtensible Markup Language) output facilitated by AI tools significantly increases efficiency, enabling the easy creation of many questions. This approach helped address common issues, such as the reuse of questions from previous years and potential copying among students.

3.1 Our Experience with AI-Assisted Quiz Design

We design quizzes incorporating a wide variety of question types to diversify the required thinking patterns of the students, to enlarge the total number of possible questions, to minimize the influence of potential communication among students:

- **Multiple Choice—Single Answer (MCSA)** is a conventional yet essential format in which only one option is correct. In Moodle, these questions use circles, with only one full inner circle marking the correct answer, known as a radio button in user interfaces.
- **Multiple Choice—Multiple Answers (MCMA):** This question type allows multiple answers to be selected. Moodle uses rectangles and check marks for the selected answers. These questions specifically test students' abilities to read all the choices and evaluate the small differences among them.
- **Computational Questions** are based on variables from a large set of possibilities and use formulas to evaluate the quality of answers. They offer diverse problem instances, ensuring different data sets for each student.
- **Short Answer Questions** require students to introduce short textual answers. This type of question is appropriate for verifying keywords in definitions, formulas, and even missing sequences of programming codes. Metacharacters are available to allow a certain degree of variation in student answers.
- **Traditional Matching:** The student gets a fixed list of items (like problems or concepts) and must select the correct corresponding answer from a dropdown menu. These questions are suitable for classification exercises, like matching the correct category to each item.
- **Drag and Drop (DnD)** interactive questions, where the students drag words or phrases into blanks inside a sentence or code. We can drag one keyword multiple times if needed. Students can also drag labels onto a background image like a diagram, a code skeleton, or a user interface. We can also implement classification problems using this technique. We use this type of question to request the students to assemble code snippets with an announced purpose by dragging keywords into predefined positions.
- **Cloze (Embedded) Questions** that combine multiple interactive questions within one question specification. They encourage comprehensive reading and analytical thinking by combining different question types.

To produce large databases of varied questions, we gave AI a proper context by prompting or providing complete documents. We also gave instructions specifying the required Moodle-compatible XML structure, the question types, and the targeted knowledge areas. This approach allowed direct import into Moodle, significantly streamlining the question generation and database population processes. For example, we used the following AI prompt to ask for a particular type of question.

Generate a Moodle XML Matching question with the title: Complexity classification.

Descriptions and correct matches:

- Array/Vector access via index: $\text{arr}[i] \rightarrow O(1)$
- The Map-Coloring problem $\rightarrow NP$
- Find an element in a set hash table;

the average case, assuming no collision. $\rightarrow O(1)$

• Finding an element in a Balanced Binary Search Tree (AVL Tree). $\rightarrow O(\log(n))$

- for (int I = 1; I < n; i++){ a[i]++; } $\rightarrow O(n)$
- Binary Search on a sorted array. $\rightarrow O(\log(n))$
- Quick Sort. $\rightarrow O(n \log(n))$
- Bubble Sort. $\rightarrow O(n^2)$

The available answers are the correct matches. An example of the question format in Moodle XML follows:

```
-<question type="matching">
-<name>
<text>Complexity classification</text>
</name>
-<questiontext format="HTML">
-<text>-<![CDATA[<p dir="ltr" style="text-align:
left;"> Classify the following problems using the
correct complexity class.</p>]]>
</text>
</questiontext>
-<generalfeedback format="html"><text/>
</generalfeedback>

<defaultgrade>0.5000000</defaultgrade>
<penalty>0.3333333</penalty><hidden>0</hidden>

<idnumber/>
<shuffleanswers>true</shuffleanswers>
-<subquestion format="HTML">
-<text>-<![CDATA[<span style="font-size: 14.9603px;">
for (int i=1;i<n;i++){a[i]++;}
</text>
-<answer><text>O(n)</text></answer>
</subquestion>
...
```

Using this technique, we generated the questions for the student survey, consulted the AI about the question types, and then designed the entire set of questions, which we present in the next section.

3.2 Students Survey

A survey with 10 questions was implemented in Moodle and administered to a group of 24 bachelor students. We categorized the questions into three groups based on the nature and purpose of the collected information:

1. Questions with quantitative answers about the preference for various working methods used for academic work preparation. These are questions 1 to 4; we present the average results in Fig. 1.
2. Questions with qualitative answers exploring students' experience and perspectives on ethical AI use: questions 5, 9 and 10.
3. The extent to which students rely on AI in different stages of academic writing. These are questions 6 to 8. We present the average results in Fig. 2.

In this survey, we used both numerical and short-answer questions. The numerical items asked respondents to type an integer between 0 and 100 to indicate how often they use or rely on a particular tool in their academic work. In this context, 100 represents a total agreement/preference. These values are not percentages but rather approximations of the frequency of the usage degree. Next, we present the questions as presented by Moodle.

As a shared aspect of all questions, we consider the context of the preparation or performing Academic Writing.

1. To what extent do you use paper and pencil?
2. To what extent do you use AI chat?
3. To what extent do you use discussions with colleagues?
4. To what extent do you use Internet search?
5. Asking an AI for collaboration while writing an essay titled "Ethics of Using AI for Writing Academic Papers", what AI would be best? What would be the prompt for asking the bot? Give a brief sample of the answer. If not, state your ideas about a paper with such a title.
6. To what extent can we use AI chat to choose a research topic?
7. To what extent can we employ AI tools for designing and writing code/programs in your academic paper without compromising ethical standards, authorship, or the ownership of your research results?
8. To what extent can we employ AI tools to write an academic paper without compromising ethical standards, authorship, or ownership of the research results?
9. What measures would be positively appreciated to encourage honesty and integrity in using AI for academic paper writing?
10. What measures would be effective in discouraging the unethical use of AI in academic paper writing?

4 Results and Discussion

The preliminary results show that students use online (AI-assisted) search, collaboration with colleagues, AI chat, and traditional methods in descending order of preference. We also observe a stronger preference for AI-assisted design and coding programs than using AI for preparation and effective writing.

Below are several answers to qualitative question 10, “What measures would be effective in discouraging the unethical use of AI in academic paper writing?”:

- Use detection tools, create clear guidelines and regulations about AI papers, foster a culture of integrity and improve transparency.
- Strict Penalties for Misconduct. Promote Ethical Scholarship.
- A multifaceted approach combining policy, education, and technological solutions is necessary to discourage the unethical use of AI in academic paper writing.

The students’ answers indicate that they intuit the limits of ethical AI use and foresee solutions to the current difficulties.

All the data for this survey is available at [19]. The repository contains the questions, the answers and the graphs with the average values for the numerical answers.

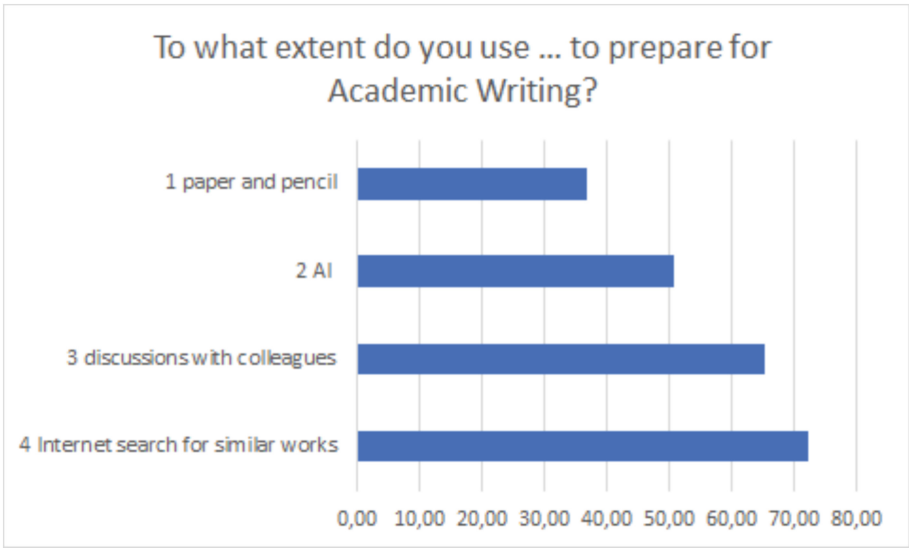


Fig. 1. Quantitative measures for work preparation.

Talking about the automatic generation of the quiz questions, not only for this particular survey but for courses where we already have an existing question database, we can take several simple actions to improve the quality of the questions. In the first step, we can export the XML of the current data set, and then we can ask an AI to check for duplicate questions and eliminate them. We can ask for a simple variation of questions. For example, for an MCMA checking for which choices are valid for a given property, say there are three answers valid from five. A similar question asking which choices **DO NOT** have the mentioned property would have two correct answers. We accentuated the syntagma “do not” to emphasize that we should also mark this way in the test to help the students who read less carefully to check for negative cases. Of course, we can ask the AI to generate similar but different questions, new questions for a new thought topic, or

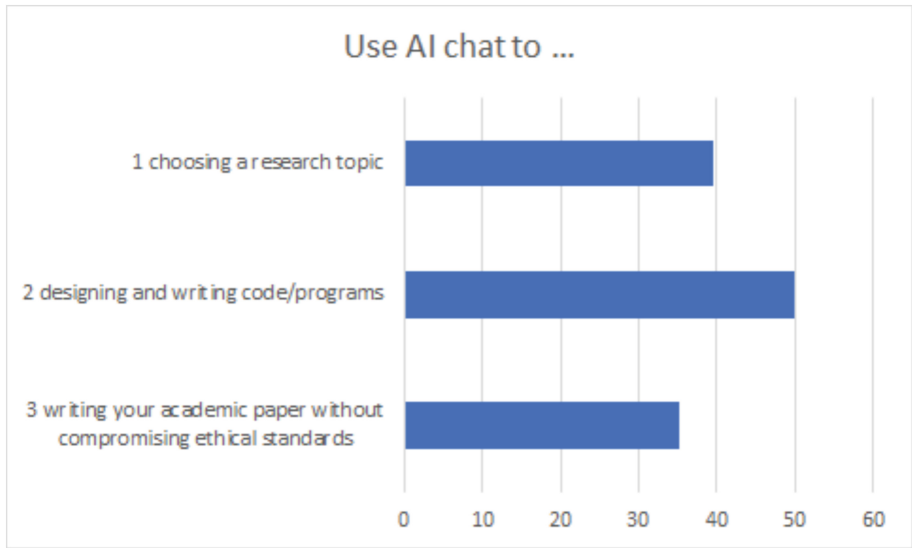


Fig. 2. The extent to which students rely on AI in different stages of academic writing.

assess the general difficulty level in filtering out the questions that are too simple or too complicated. After importing the new questions, we should check the quality of the AI reasoning for accuracy, correctly marked answers, and ambiguity or omissions.

We observed an interesting aspect of learning programming languages or other coding techniques using this online test. Some students performed well during tests, which contained MCMA, classifications, and drag and drop, but when asked to write a new program from scratch, they could not start. In our interpretation, there are different abilities: connecting to an existing code, understanding something written already, even observing inconsistencies on one side, composing a new structure, and writing something on a blank page, which are different categories of skills. To compensate for this student's difficulty, we should encourage them to participate in individual small, full-scale projects as homework or during class. Our observation is in concordance with the previous knowledge from the literature [20], specifically Raymond Lister's two "natural" stages of cognitive development for programmers:

- *Preoperational Reasoning* includes programmers' capabilities required for tracing/understanding the functionality of a code fragment without realizing an abstract meaning.
- *Formal Operational Reasoning* includes capabilities of code abstraction, processing diagrams, and design specifications.

MCMA, short answers, and DnD or Cloze questions test students' preoperational reasoning capabilities. Full-scale programming tasks, team projects, and the effort to write about their work could help the students acquire a new cognitive capacity related to formal operational reasoning. We assist them; they also have the new AI questioning

capabilities to solve their difficulties. However, they must still work individually or exercise group communication to reach the new cognitive level.

5 Concluding Remarks

Using AI opens new opportunities for enhancing academic writing productivity; however, we must consider the risk levels associated with this use.

This article revises various documents related to the applications of generative AI tools in HEIs. We present a supervised method to enhance the question database for online tests by adding various types of questions using AI. With the help of this method, we generated a test questionnaire for students to test their readiness for using AI tools for their work.

Our questionnaire reveals that students perceive AI as applicable but with well-defined limits. It is widely accepted for literature review and code-generation tasks, but could compromise academic integrity if used inappropriately in the final writing process. Even with a small sample of students participating in our questionnaire, the results provide us with valuable insights. Our data reveal a clear preference of students for AI-assisted web search and peer collaboration, while traditional paper-and-pencil methods trail far behind. This pattern supports the idea that Gen-Z learners already treat AI as an extension of online search rather than a separate, exotic aid.

Qualitative answers cluster around three main themes: institutional policy, education for integrity, and technical detection, indicating that students intuit a balanced approach toward the future AI-assisted practices, rather than banning them.

The results also suggest that ethical AI usage should focus on promoting responsible integration, providing explanations, and offering examples of practices that could result in potential risks.

In a future paper, we will share our experience with the methodology for producing course content slides using AI for theoretical support and laboratory work for programming courses.

Acknowledgements. This text started with an intensive series of conversations with ChatGPT [21] models GPT 4o, o1, o3, o4-mini and o4-mini-high, helping as collaborative work during various stages of preparing the current material: designing the questions for the student questionnaire, discussing the structure of this document, and giving “reviewer-like comments”. However, the authors assume the responsibility for the textual content of this article.

After the manuscript draft, Grammarly refined the original text for better readability and style [22].

We thank Prof. Claudio Moraga for his kind observations on our manuscript.

References

1. Yeo, M.: Academic integrity in the age of Artificial Intelligence (AI) authoring apps. *TESOL* **14**, 03 (2023)
2. “IEEE Code of Ethics,” [Online]. Available: <https://www.ieee.org/about/corporate/governance/p7-8.html>

3. "IEEE Code of Conduct," [Online]. Available: https://www.ieee.org/content/dam/ieee-org/ieee/web/org/about/ieee_code_of_conduct.pdf
4. "ACM Code of Ethics and Professional Conduct," [Online]. Available: <https://www.acm.org/code-of-ethics>
5. "Springer Nature, Editorial Policies: Artificial Intelligence (AI)," [Online]. Available: <https://www.springer.com/us/editorial-policies/artificial-intelligence--ai-/25428500>
6. "Generation Z," Merriam-Webster, [Online]. Available: <https://www.merriam-webster.com/dictionary/Generation%20Z>
7. "AI maturity toolkit for tertiary education," Joint Information Systems Committee (JISC), [Online]. Available: <https://www.jisc.ac.uk/ai-maturity-toolkit-for-tertiary-education>
8. Webb, M.: Generative AI - a primer, report v 1.4. JISC (2024). <https://repository.jisc.ac.uk/9714/1/generative-ai-primer-aug-2024.pdf>
9. Attewell, S.: Embracing Generative AI in Assessments: A Guided Approach. (2024). <https://nationalcentreforai.jiscinvolve.org/wp/2024/08/14/embracing-generative-ai-in-assessments-a-guided-approach/>
10. Grandinetti, P.: Strategies and Tools for Academic Integrity in Online Environments | Teaching and Learning Resource Center. <https://teaching.resources.osu.edu/teaching-topics/strategies-tools-academic-integrity>
11. Knight, S., Walker, S.: Rethinking Assessment and Feedback – Shifting to Digital Assessment, 36 min. (2022). <https://jiscpodcast.libsyn.com/rethinking-assessment-and-feedback-shifting-to-digital-assessment>
12. Sabzalieva, E.: ChatGPT and artificial intelligence in higher education: quick start guide (2023). <https://unesdoc.unesco.org/ark:/48223/pf0000385146>
13. Miao, F.H.W.: Guidance for generative AI in education and research, UNESCO, p. 44 (2023)
14. Opportunities, Guidelines and Guardrails on Effective and Equitable Use of AI in Education. OECD Publishing (2023). https://www.oecd.org/content/dam/oecd/en/about/projects/edu/smart-data-and-digital-technology-in-education/Opportunities,%20guidelines%20and%20guardrails%20for%20effective%20and%20equitable%20use%20of%20AI%20in%20education.pdf/_jcr_content/renditions/original./Opportunities,%20guidelines%20and%20guardrails%20for%20effective%20and%20equitable%20use%20of%20AI%20in%20education.pdf
15. International Summit on the Teaching Profession (2023)
16. High-level summary of the AI Act (2024). <https://artificialintelligenceact.eu/high-level-summary/>
17. [Online]. Available: <https://artificialintelligenceact.eu/>
18. Attewell, S.: Exploring AI and assessment – avoid, outrun or embrace (2024). <https://www.jisc.ac.uk/blog/exploring-ai-and-assessment-avoid-outrun-or-embrace>
19. Dediu, A.-H.: [Online]. Available (2024). <https://github.com/ahdediu/AI-Writing-Survey.git>
20. Lister, R.: Concrete and other neo-piagetian forms of reasoning in the novice programmer. In: Proceedings of the Thirteenth Australasian Computing Education Conference (ACE 2011). Perth, Australia, Perth (2011)
21. "OpenAI" (2023). <https://chat.openai.com/>
22. "Grammarly" (2024). <https://www.grammarly.com/>



Towards a Gender-Inclusive Tech Landscape in Portugal: Women4Digital's Insights on Gender and Digital Transformation

Rosa Monteiro¹ , Mariana Santos² , and Lina Coelho¹ 

¹ Faculty of Economics, University of Coimbra and Centre for Social Studies of the University of Coimbra, Coimbra, Portugal

{rosamonteiro,women4digital}@ces.uc.pt, lcoelho@fe.uc.pt

² Centre for Social Studies of the University of Coimbra and Faculty of Economics of the University of Coimbra, Coimbra, Portugal
marianasantos@ces.uc.pt

Abstract. This paper presents preliminary findings from the Women4Digital project, which analyze public policies in Portugal aimed at promoting women's participation in the Information and Communication Technologies (ICT) sector. Despite national initiatives such as Engineers for a Day and the integration of gender equality goals into broader strategies like ENIND and INCoDe.2030, women remain significantly underrepresented in ICT careers and leadership positions. In this article, we present some results from the diagnosis of the situation in Europe and Portugal, the normative European framework and some analysis of our mapping of national policies to promote gender equality in the digital in Portugal. The study highlights the need for systemic and intersectional political approaches, with the full integration of a gender intersectional perspective, because the inequality problem is not only a matter of individual-level interventions but implies profound systemic gender transformations in education and labour market systems.

Keywords: Gender Equality · Digital Transformation · STEM · Public Policy

1 Introduction

This article presents the preliminary findings of the Women4Digital project, which was designed to analyse and improve the public policies of gender mainstreaming, in Portugal, that promote women's participation in the ICT sector. The gender gap in the ICT sector remains a significant challenge, with women accounting for only 19.4% of ICT specialists employed in the EU, compared to 80.6% of men [15]. This disparity not only limits women's opportunities within the industry but also hinders the potential for an inclusive and egalitarian digital transformation across all sectors of society.

This is a problem that is accentuated in the ICT sector and professions but presents itself more broadly in what is usually grouped under the acronym STEM. The acronym STEM includes sciences, technologies, engineering, and mathematics, but as an umbrella term used to group together disciplines that can be very different, even when we look

for gender gaps in these educational and professional realms. For instance, there is a globally bigger representation of women in sciences and mathematics than in engineering and technologies. Alice Eagly [6] proposes the concept of “inconsistent gender gap” to express this differentiation. In fact, according to Eurostat data, while women represented around 51% of students in natural sciences, mathematics and statistics; only 27% of students in engineering, manufacturing and construction, and around 20% of students in ICT were female [6].

The phenomenon concerns horizontal segregation, which manifests both within education, through gendered choices of study fields, and within the labour market, through the unequal distribution of men and women across occupational sectors. As demonstrated by Coelho and Ferreira [4], despite the growing participation of women in higher education and employment, occupational structures remain deeply gendered. Several factors contribute to the persistence of horizontal segregation in educational choices, which later impacts the labour market by clustering men and women into segregated professions and sectors. Factors such as gender stereotypes, STEM learning experiences, individual background, contextual influences – including parental and peer support, role models, societal and cultural factors – as well as school and national education policies, play significant roles in shaping girls’ career aspirations.

The issue of gender inequality in education, particularly in STEM fields, has progressively occupied a prominent place on the agenda of international organizations. Various international regulatory frameworks have been established to promote and reinforce recommendations for gender-inclusive education, such as the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), the Beijing Platform for Action, and more recent European initiatives and strategies. Specifically, the European Education Area (EEA) sets out objectives for national reforms and European cooperation in education and training, including: (1) fostering greater gender sensitivity within educational processes and institutions; (2) actively challenging and dismantling gender stereotypes; and (3) promoting a proper gender balance in leadership positions, particularly within higher education institutions [30].

In this context, the European Union (EU) has recently strengthened its commitment to gender-sensitive STEM education, setting concrete targets to enhance the participation of girls and women in STEM and technological fields. The Union of Skills roadmap outlines a comprehensive strategy aimed at ensuring equal access to digital education, including the creation of a robust EU digital education ecosystem. To address the decline in STEM performance and improve student engagement, the European Commission has proposed a Strategic Plan for STEM Education (COM(2025) 89 final) with a central objective of boosting female participation. By 2030, the targets include having at least 45% of students in medium-level vocational education and training (VET) enrolled in STEM fields, of whom at least 25% should be women; achieving 32% of tertiary-level students in STEM, with a minimum of 40% women; and ensuring that students enrolled in ICT PhD programs represent at least 5% of all PhD students, with at least one-third of those in ICT PhD programs being women. These ambitions are part of broader efforts to enhance digital competences, with the Digital Competence Framework scheduled for revision by 2025 to incorporate emerging technologies such as artificial intelligence, thereby fostering a future-oriented digital education system in the EU [7].

Nevertheless, despite the existence of clear strategic frameworks and ambitious objectives, the actual implementation of these recommendations often remains partial and inconsistent in Portugal, as highlighted by recent analyses [17, 30].

This article presents some results from the ongoing work of the project Women4Digital – Gender Equality in the Digital Transition in Portugal: What Place for Women?, financed by the Portuguese Foundation for Science and Technology. The project seeks to address this imbalance by analysing the national policies designed to combat horizontal gender segregation and encourage more inclusive participation in ICT careers, with a focus on understanding and monitoring the impact of these policies since 2017. This project aims to analyse how Portuguese policies have attempted to challenge these structural problems, related stereotypes and increase female participation in the digital sector, examining both successes and ongoing obstacles.

2 Gender Segregation in ICT, a Persistent Problem

The gender gap in ICT is not confined to Portugal; it is a widespread phenomenon that stems from deep-rooted structural gender inequalities in education, labour markets, and cultural expectations including gender equality paradox influencing occupational segregation phenomena [2]. The underrepresentation of women in ICT careers begins at an early stage, often influenced by stereotypes that associate technology and engineering fields with men. These biases discourage young girls from pursuing studies in science, technology, engineering, and mathematics (STEM), resulting in fewer women entering ICT-related careers.

Cross-cultural studies reveal that women's underrepresentation in computing is not a universal phenomenon, with experiences varying significantly depending on context [18]. Furthermore, there is no single dominant cause for this disparity [21]. Several factors have been discussed in the literature regarding how the gender gap in ICT and the broader technology field was created, stemming from a complex set of factors that operate at the individual, organizational, and systemic levels. At the individual level, aspects such as low self-confidence [10, 40], lack of identification with science [5, 41] and perceived lack of aptitude or interest [25] play a significant role. Social and familial expectations also influence girls' career paths, often discouraging them from pursuing STEM fields [11, 26]. At the organizational level, college experience, lack of support structures, and faculty biases [29] can negatively affect women's progression and sense of belonging. Stereotypes embedded in educational systems – such as the idea that brilliance or genius is a male trait – also contribute to early disidentification from STEM [1]. At the systemic level, broader cultural and gender norms – including the widespread stereotype that STEM is a male domain – continue to shape perceptions and opportunities [30]. Additionally, labour market factors such as the masculinization of the sector, hostile or exclusionary work environments, and the lack of female role models further hinder women's participation and retention in ICT [30], compounded by workplace cultures that reinforce traditional gender norms, inflexible work hours, and the burden of unpaid care responsibilities [23] these structural barriers collectively contribute to attrition among women professionals in tech.

Fagenson [16] presents three complementary perspectives that seek to explain the underrepresentation of women: individual, cultural, and structural or institutional. The

individual perspective, often referred to as the “fix the women” approach, focuses on mentoring, training, and coaching, and attributes underrepresentation to psychological, personality, socialization, or career orientation differences between women and men. The cultural perspective, on the other hand, suggests that the lack of women is linked to the history, culture, and policies of the organization, as well as broader societal cultural factors that influence these dynamics. Finally, the structural perspective highlights how organizational hierarchies, procedures, and both formal and informal rules that govern the functioning of institutions can hinder the recruitment and promotion of women.

Taking all these interconnected factors and causes into account, it becomes clear that gender inequalities in ICT is not simply a matter of addressing a single cause but rather recognizing the interplay of various elements that shape the dynamics of inequality. Therefore, effective interventions must be multifaceted and context-sensitive, considering both the structural and cultural dimensions at play [21].

3 Combat Strategies: Insights from the Literature on Successes and Setbacks

Building on these insights, it is essential to examine strategies for addressing gender inequality in the digital sphere, particularly through public policies that promote gender equality in digital environments. The literature identifies several key success factors. First, programs should implement a comprehensive set of measures rather than isolated or ad hoc interventions [24, 27] adopting a multilevel and multi-actor approach [29] and addressing three core dimensions of digital inclusion: access, motivation, and capability [24]. These initiatives must ensure continuity over time [24, 27] and go beyond individual-focused actions, such as peer mentoring or social events [24]. Moreover, policies must tackle the structural and cultural barriers that sustain gender inequalities in science and technology, thereby challenging “gender-biased universalism” and the myth of gender neutrality [13, 27]. Effective interventions must be institutionally embedded – whether in public policy, companies, schools, or universities [24] – and grounded in feminist and gender-based analysis [34]. Successful strategies also depend on broad networks and external partnerships, strong feminist alliances across civil society and institutions, and robust evaluation frameworks that measure not only implementation and outputs but also transformative outcomes [27, 28]. Importantly, assessment metrics must consider structural challenges such as online harassment, gendered workplace norms in tech, and the sexual division of labour in digital sectors [35]. A strong communication strategy and sustained political commitment – especially from high-level decision-makers in sectors beyond gender equality, including digitalization, education, and modernization – are also critical [24]. In our study, we draw on these success factors to analyze Portuguese public policies implemented since 2017 aimed at promoting women’s participation and addressing horizontal gender segregation in the digital domain.

We draw on McKinnon’s [27] key critique of the “delivery emphasis” in gender equality policies in science and technology – namely, the focus on quantifiable outputs (e.g., participant numbers or satisfaction) over deeper, long-term transformations in institutional culture, practices, or power structures. As McKinnon [27] highlights, such narrow evaluation frameworks can mask whether initiatives truly challenge systemic

inequalities. Given that individuals or isolated organizations cannot alone shift institutional norms, it is crucial to move beyond fragmented, short-term actions. Although useful in specific cases, these limited interventions often lack the coherence and ambition necessary for structural change, leading to superficial adjustments or tokenism [27].

Our study thus explores how effective gender equality in the digital sector requires coordinated, multi-level, intersectional, and sustained strategies rooted in shared visions and long-term structural reform. In Portuguese context, some recent studies analyzed policy implementation addressing gender inequalities in ICT and STEM in secondary education [30, 39] highlighting a significant lack of structured actions, contrasting sharply with national and international policy commitments. They further emphasize the absence of systemic, sustained interventions and the limited integration of gender mainstreaming in public projects related to the digital transformation of education, exposing a persistent gap between policy ambitions and practical implementation. Those studies identify a key obstacle to implementing gender equality initiatives in education: the very absence of projects designed with a gender-sensitive framework. While some initiatives promote youth participation in STEM, they largely overlook the specific underrepresentation of girls, failing to adopt targeted measures or gender mainstreaming strategies.

4 Materials and Methods

This article presents some results from the ongoing work of the project Women4Digital – Gender Equality in the Digital Transition in Portugal: What Place for Women?

The research is based on a desk survey and document analysis, in interviews and focus group. To undertake a diagnostic analysis of the situation of women and gender equality in the digital world, we performed a comprehensive statistical analysis of relevant data sources (She Figs. 2024, Women in the Digital Age, Union of Skills, and the Portugal 2024 Digital Decade Country Report). The desk survey aims at the systematic mapping and review of the existing public policies in Portugal to promote women's participation in the digital and the relevant European regulations.

In this article, we present the preliminary results of these first two components of the study, namely the diagnostic analysis of the current situation and some preliminary results of the mapping of relevant regulations and public policies.

5 Results

5.1 Gender Disparities in the Digital: A Preliminary Diagnosis

Between 2013 and 2023, the ICT sector expanded approximately three times faster than the total economy in OECD countries, drastically reshaping labour markets and economic structures [31]. However, this accelerated digital transformation has not translated into gender inclusivity in ICT domains like employment and self-employment, higher education and doctorates, patents, and start-ups. As of 2023, women represented merely 31% of the workforce in the ICT sector across the OECD, with male ICT specialists outnumbering their female counterparts by a factor of three to eight. Notably, the share of women in the field has increased by only one percentage point over the last decade [31].

Data from the European Union reflect a similar scenario: in 2023, only 19.4% of ICT specialists were women, although the number of female ICT specialists grew by 85.6% between 2013 and 2023, compared to 53.9% for men [15]. Women are significantly overrepresented in care and human-centered roles – sectors typically associated with lower pay compared to STEM occupations – and are less likely to occupy high-paying roles within tech industries. In 2021, only 22.7% of ICT degree holders were women, a pattern mirrored in vocational education and training [32]. Moreover, even women with STEM qualifications often do not enter or remain in ICT careers (Table 1).

Table 1. Gender Disparities in ICT.

Indicator	Value	Source
Women in the ICT workforce across the OECD	31%	[20]
Women among ICT degree holders	22.7%	[15]
Women as inventors in ICT-related patents	4%	[15]
Women in ICT patent co-authorship	20%	[32]
Women among the self-employed in ICT	18%	[8]
Women with employees among self-employed in ICT	12%	[8]
Start-ups in ICT with only female founders (venture-capital funded, 2000–2021)	6%	[32]
Women completing doctorates in ICT (EU-27)	1.5% (4.1% men)	[32]
Women completing doctorates in ICT in Portugal	0.9% (4.1% men)	[32]
Women among ICT specialists in the EU	19.4%	[14]
Growth in number of female ICT specialists (EU, 2013–2023)	+ 85.6%	[14]

Source: Own elaboration

Besides labour market numbers, education plays a major part in gender segregation in ICT. In fact, gender disparities begin early. PISA 2022 data reveal that career aspirations diverge sharply: less than 1.5% of girls across the OECD express a desire to work in ICT, compared to almost 10% of boys. This gender gap persists even among the highest-performing students in math and science [33]. Among individuals aged 16–24, the ability to program – a critical digital skill – shows significant gender disparity, disadvantaging young women from early career stages.

Portugal registered significant declines between 2018 and 2021 in the proportion of female doctoral graduates in key STEM areas, particularly Mathematics and Statistics (–22.8 P.P.), Physical Sciences (–16.2 P.P.), and ICT (–7.1 P.P.), diverging from the modest upward trends observed at the EU level [7]. Only 19.2% of doctoral graduates in ICT were women, between 2021 and 2022 [8].

While some fields, such as Engineering and Engineering Trades (+7 P.P.), showed improvement, the overall trajectory signals a concerning backslide in women's participation in advanced STEM education. Recent studies also suggest a growing ideological

divide regarding gender equality, particularly among younger generations in OECD countries [19, 22, 42].

5.2 Mapping European Regulatory and Normative Context

Given the prominent role played by the EU in promoting equality policies – often acting ahead of its Member States it would be pertinent to analyse in greater depth the processes and impacts of this dynamic of Europeanization [36]. Since the 1990, European normative frameworks have integrated gender concerns into STEM and digital policies, emerging from the major concern of the underrepresentation of women in STEM fields in Europe and to halt the “leaky pipeline” effect [3, 30, 37].

Nowadays, the EU’s policy efforts towards gender equality in the digital sector are integrated into the Digital Decade Policy Programme 2030 (Decision (EU) 2022/2481). Where it is approached through the concept of “gender balance,” (Article 3). The Programme reveals a particular emphasis on education, training, and the importance of gender-disaggregated data to monitor progress, and the guarantee of participation of women’s representatives. Concrete targets for the promotion of gender equality are detailed across several annexes of the Programme, aiming for instance, the gender gap in basic digital skills and the participation of women in ICT professions. In Annex 3 Portugal is noted for having strong potential for achieving an inclusive digital transformation but is advised to sustain and intensify efforts to enhance the digital skills of its population and to train more ICT specialists, including women [7].

The monitoring of the EU’s digital transformation is grounded in systematic and evidence-based tools. Since 2014, the Digital Economy and Society Index (DESI) has been the primary framework for assessing the digital performance of EU Member States. DESI evaluates multiple aspects of digitalization, including digital skills, connectivity, the integration of digital technologies in businesses, and the digitalization of public services. As part of the EU’s Digital Decade Policy Programme 2030, DESI has undergone significant restructuring, with the 2023 edition fully integrated into the Digital Decade Report on the State of the Digital Decade. This integration strengthens the EU’s ability to track progress towards the 2030 digital targets, while fostering a more holistic approach to monitoring the digital transformation across four key dimensions: Digital Skills; Digital Infrastructure; Digital Transformation of Businesses; Digitalization of Public Services.

While DESI tracks progress across several dimensions, it also highlights the persistent gender gap in the digital economy, particularly in areas like ICT employment, digital skills, and entrepreneurship. To address this, the Women in Digital (WiD) Scoreboard was introduced as part of the DESI framework. This tool focuses on monitoring and measuring women’s participation in the digital sector, identifying barriers to digital inclusion and promoting gender equality across the EU.

The WiD Scoreboard evaluates Member States’ performance through 12 indicators, grouped into three key dimensions: Internet Use, Internet User Skills, and Specialist Skills and Employment. By including gender-focused indicators, the WiD Scoreboard contributes to a more inclusive monitoring process, ensuring that women’s roles in the digital economy are given the attention they deserve.

The recent DESI 2023 data reveal significant disparities in digital performance across Member States, particularly in relation to gender equality. These disparities underscore the importance of targeted interventions, especially in areas where women remain under-represented in digital fields or lack access to the necessary skills and infrastructure to succeed in the digital economy.

While the WiD Scoreboard represents a relevant effort to monitor women's participation in the European digital economy, several critiques can be raised when this tool is examined through the lens of feminist digital scholarship. As Prešički [35] argues, there is a structural tendency to silence women's experiences in the digital age – not only through technological practices but also via public policies that fail to account for intersectional gender dynamics. The WiD Scoreboard, by relying exclusively on quantitative indicators such as the number of ICT specialists or levels of internet use, risks reducing women's digital participation to decontextualized metrics, overlooking structural barriers such as online harassment, gendered norms in tech workplaces, and the sexual division of labour in digital sectors.

Education and training play a central role in shaping access to digital opportunities and are thus critical areas for promoting inclusion. In response to the urgent need for digital transformation in education, the EU has launched a series of initiatives seeking to address the structural inequalities that digitalisation has both revealed and intensified, particularly during the COVID-19 pandemic [12].

Digital education policy is now embedded within the broader EU vision for a green and digital transformation. The Digital Education Action Plan 2021–2027 (COM(2020) 624 final) contributes to building a Europe fit for the Digital Age and supports the objectives of NextGenerationEU, the European Skills Agenda, the European Pillar of Social Rights Action Plan, and the Digital Compass 2030.

Gender equality is integrated throughout these efforts. The European Education Area (EEA) establishes as key priorities the development of gender-sensitive education systems, the dismantling of gender stereotypes, and the promotion of balanced gender representation in leadership roles, including within higher education institutions [30]. Recent EU commitments also include setting concrete targets to increase the participation of girls and women in STEM and ICT fields. The Union of Skills roadmap (COM(2025) 90 final) outlines a comprehensive strategy for ensuring equal access to digital education, emphasising the creation of an inclusive and future-oriented education ecosystem. To address declining STEM performance and boost engagement, the EU has proposed a STEM Education Strategic Plan (COM(2025) 89 final) that explicitly aims to increase female participation. By 2030, targets include ensuring that at least 45% of students in medium-level vocational education and training (VET) are enrolled in STEM fields, with a minimum of 25% being female; at least 32% of tertiary-level STEM students, with a minimum of 40% women; and at least 5% of students in ICT doctoral programmes, of whom at least one-third should be female [9]. Furthermore, Horizon Europe funds consortia of universities and research centres to develop gender equality plans in higher education institutions, making their existence a condition for funding [30].

As digital transformation continues to reshape the European labour market, the EU has prioritised the integration of gender equality across its digital and employment strategies.

This endeavour is achieved in the Gender Equality Strategy 2020–2025 which explicitly acknowledges the structural underrepresentation of women in ICT and technology-driven industries and supports targeted initiatives to redress this imbalance through upskilling, education, and access to opportunities. The strategy is further supported by legal instruments such as the Work-Life Balance Directive (Directive (EU) 2019/1158), and the Pay Transparency Directive (Directive (EU) 2023/970), and the Directive on Gender Balance on Corporate Boards (Directive (EU) 2022/2381).

Complementary policy frameworks such as the Digital Education Action Plan (COM(2020) 624 final), the European Skills Agenda (COM(2020) 274 final), and the Digital Compass 2030 COM(2021)118 final reinforce these objectives by promoting inclusive learning and training ecosystems.

Within the Recovery and Resilience Facility (RRF), one of the central pillars of NextGenerationEU, National recovery plans are expected to embed gender mainstreaming into all reform areas, including labour market policies, the digitalisation of public services and SMEs, and the development of inclusive digital infrastructures.

6 Mapping Public Policy Initiatives for Gender Equality in the Digital Transformation in Portugal

The political agenda to promote gender equality in Portugal's digital sector was initiated in 2017 with the launch of the Engineers for a Day programme, under the governmental portfolio for Equality [29]. While the process of Europeanization of equality public policies had already been underway for several decades [30], the gender mainstreaming agenda within the digital sector only gained momentum with this political initiative of positive action, in 2017, and with the gender mainstreaming approach in the National Strategy for Equality and Non-discrimination (ENIND) 2028–2030. In fact, in our analysis we've concluded that the government, through the gender equality tutelage has been the main actor to foster the gender equality perspective in digital agenda.

Between 2017 and 2022, a total of 32 political initiatives were identified, comprising 15 measures directly addressing the digital sector and 17 with an indirect focus, alongside seven major public policy programmes. In terms of overarching political strategies and action plans, the integration of gender equality in the field of ICT is explicitly addressed in a gender mainstreaming policy in the National Strategy for Equality and Non-Discrimination 2018–2022 (ENIND) – the first national strategy to incorporate this theme – which has since been maintained across successive frameworks. Relevant strategic instruments also include INCoDe.2030, the National Strategy for Artificial Intelligence (2019), the Action Plan for Digital Transition (2020), the Recovery and Resilience Plan (2021–26) (notably its Digital Transition dimension 2.3), the Public Investment Programme in Research and Development (2021–2030), the Alliance for Equality in ICT (2021), the Pessoas 2030 Programme (2023), and the National Digital Strategy (2024). The analysis examined the integration of gender considerations into national digital strategies and related policy frameworks, with particular emphasis on the interplay between gender equality and digital agendas, as well as the institutional actors involved in their design and implementation. To offer a comprehensive overview

of gender mainstreaming in this context, the scope extended beyond explicitly gender-focused digital policies to include relevant measures within broader sectors such as education, science and technology, economy, structural funds, labour, and vocational training. We also assessed whether the initiatives had a continuous or ad hoc character, and whether they aimed to transform structural inequalities or simply “fix the women,” focus on adapting women to pre-existing structures rather than changing those structures themselves.

Currently, gender equality objectives and concrete measures are incorporated into the National Strategy for Equality and Non-Discrimination (ENIND) 2023–2030 and the National Digital Strategy (NDS) (2024), particularly through its Action Plan 2025–2026. This NDS strategy frames the issue primarily as a business case, emphasising the economic and individual impacts of women’s underrepresentation in the digital. It attributes the causes mainly to gender stereotypes in education and the labour market, as well as insufficient support for women pursuing digital studies and careers. While it sets out to create more inclusive conditions to retain women in digital fields, it falls short of addressing the structural and intersectional barriers underlying the problem. The strategy aligns with the targets of the Digital Decade 2030 and the Digital Economy and Society Index (DESI) [38]. To operationalise this strategy, the Action Plan 2025–2026 outlines “Initiative #2 – the National Girls in STEM Programme”, structured around three main actions: 2.1 – Establishment of the Programme’s Network of Ambassadors; 2.2 – Creation of the “Girls in STEM” Emblem; and 2.3 – Implementation of the National Programme for Girls in STEM. Although these actions were scheduled for the first half of 2025, no concrete developments have been reported to date. The initiative identifies three priority areas – education, higher education, and the labour market – and foresees the involvement of several institutional actors, from modernization, equality, education and digital inclusion sector and higher education institutions (IES).

7 Final Remarks

In conclusion, the ongoing research of the Women4Digital project highlights the essential role of the National Strategy for Equality and Non-Discrimination (ENIND) in introducing gender equality into public policies, particularly within the digital transformation in Portugal. The ENIND framework, with its intersectional, systematic and multi-level and multi-actor approach, robust gender diagnosis, and strong political commitment, has been a key political mechanism for advancing gender equality across various sectors. The National Girls in STEM Programme, as outlined in the Action Plan 2025–2026 of the National Digital Strategy (2024), is an example of the kind of initiatives that stemmed from the ENIND’s vision, aiming to reduce gender disparities in STEM education and careers.

However, our research points to a critical gap in the new digital strategy’s approach. While the National Digital Strategy continues to address gender equality within the digital, it predominantly focuses on economic arguments and individual-level causes, neglecting the more comprehensive, structural, and intersectional framework that was introduced by the ENIND. The new strategy’s focus on business-centric arguments, focusing more on the supply side than on the demand side, risks undermining the broader

social and political goals of gender equality, thus limiting its potential to sustain and build upon the progress made. Our first results point to a deceleration in political voices on this agenda, on the systemic and feminist framing of the problem, and a tendency to a more technocratic orientation, in a strategy which privilege certain elite groups of women rather than an intersectional and inclusive one, risking reinforcing the very structures they seek to challenge, offering visibility without transformation [27]. As the research progresses, further insights will emerge regarding the effectiveness of the current policies and the potential for aligning them more closely with the gender equality goals established in previous frameworks, such as the ENIND.

Acknowledgments. This study was funded by PlanAPP and the Foundation for Science and Technology (FCT) through the Science4Policy (S4P) Programme: Public Policy Science Studies.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Archer, L., Dewitt, J., Osborne, J., Dillon, J., Willis, B., Wong, B.: Not girly, not sexy, not glamorous”: Primary school girls’ and parents’ constructions of science aspirations. *Pedagog. Cult. Soc.* **21**(1), 171–194 (2013). <https://doi.org/10.1080/14681366.2012.748676>
2. Breda, T., Jouini, E., Napp, C., Thebault, G.: Gender stereotypes can explain the gender-equality paradox. *Proc. Natl. Acad. Sci.* **117**(49), 31063–31069 (2020). <https://doi.org/10.1073/pnas.2008704117>
3. Cavaghan, R.: Making gender equality happen: Knowledge, change and resistance in EU gender mainstreaming. Routledge (2017)
4. Coelho, L., Ferreira, V.: Segregação sexual do emprego em Portugal no último quarto de século – Agravamento ou abrandamento? *e-cadernos CES* (29), 77–98 (2018). <https://doi.org/10.4000/eces.3205>
5. Diekman, A.B., Brown, E.R., Johnston, A.M., Clark, E.K.: Seeking congruity between goals and roles: a new look at why women opt out of science, technology, engineering, and mathematics careers. *Psychol. Sci.* **21**(8), 1051–1057 (2010). <https://doi.org/10.1177/0956797610377342>
6. Eagly, A.H.: Hidden in plain sight: the inconsistent gender gaps in STEM and leadership. *Psychol. Inq.* **32**(2), 89–95 (2021). <https://doi.org/10.1080/1047840X.2021.1930764>
7. European Commission: Comunicação da Comissão ao Parlamento Europeu, ao Conselho, ao Comité Económico e Social Europeu e ao Comité das Regiões – Estado da Década Digital 2024 (2024). <https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A52024DC0274>
8. European Commission: She Figures 2024: Gender in research and innovation – Statistics and indicators. Publications Office of the European Union (2025a). <https://op.europa.eu/en/publication-detail/-/publication/7646222f-e82b-11ef-b5e9-01aa75ed71a1/language-en>
9. European Commission: A STEM education strategic plan: Skills for competitiveness and innovation (COM(2025) 89 final) (2025b). https://education.ec.europa.eu/sites/default/files/2025-03/STEM_Education_Strategic_Plan_COM_2025_89_1_EN_0.pdf
10. Eccles, J.S.: Gendered educational and occupational choices: Applying the Eccles et al. model of achievement-related choices. *Int. J. Behav. Develop.* **35**(3), 195–201 (2011). <https://doi.org/10.1177/0165025411398185>

11. Eccles, J.S., Jacobs, J.E., Harold, R.D.: Parents and gender-typed socialization during the middle childhood and adolescent years. In: Beall, A.E., Sternberg, R.J. (eds.) *The psychology of gender*, pp. 127–153. Guilford Press (1993)
12. European Institute for Gender Equality (EIGE): Gender Equality Index 2020: Digitalisation and the future of work. Publications Office of the European Union (2020). https://eige.europa.eu/publications-resources/toolkits-guides/gender-equality-index-2020-report?language_content_entity=en
13. Engeli, I., Mazur, A.: Research on French gender equality policies in practice: One step closer to a recipe for success. *LIEPP Policy Brief* (2021). https://www.sciencespo.fr/liepp/sites/sciencespo.fr/liepp/files/LIEPP_PB_55.pdf
14. Eurostat: Self-employment by sex, age and occupation (1 000). Employment and Unemployment (LFS) Database (2024). https://ec.europa.eu/eurostat/databrowser/view/lfsa_esgan2_custom_7960655/default/table?lang=en
15. Eurostat: ICT specialists in employment. *Statistics Explained* (2025). Retrieved 8 May 2025, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=ICT_specialists_in_employment
16. Fagenson, E.A.: At the heart of women in management research: Theoretical and methodological approaches and their biases. *J. Bus. Ethics* **9**(4), 267–274 (1990). <https://doi.org/10.1007/BF00380326>
17. Ferreira, V., Lopes, M., Vieira, C.C., Monteiro, R., Santos, C.C.: Integração dos estudos de género nos currícula e práticas pedagógicas no ensino superior em Portugal. *Centro de Estudos Sociais, Coimbra* (2024). <https://www.ces.uc.pt/pt/investigacao/projetos-de-investigacao/projetos-financiados/engender>
18. Frieze, C., Quesenberry, J.L., (eds.): *Cracking the digital ceiling: Women in computing around the world*. Cambridge University Press (2019)
19. Government of France: Rapport annuel 2024 sur l'état des lieux du sexisme en France: S'attaquer aux racines du sexisme. Haut Conseil à l'Égalité entre les femmes et les hommes (2024). https://www.haut-conseil-egalite.gouv.fr/IMG/pdf/hce_-_rapport_annuel_2024_sur_l_etat_du_sexisme_en_france.pdf
20. International Labour Organization (ILO): Employment by sex and economic activity (thousands) – Annual. *ILOSTAT Statistics on Employment* (2023). <https://ilostat.ilo.org/topics/employment/>
21. Kalpazidou Schmidt, E., Cacace, M.: Setting up a dynamic framework to activate gender equality structural transformation in research organizations. *Science and Public Policy* **46**(3), 321–338 (2019). <https://doi.org/10.1093/scipol/scy059>
22. King's College London, Ipsos: Emerging tensions? How younger generations are dividing on masculinity and gender equality. *The Global Institute for Women's Leadership & The Policy Institute* (2024). <https://www.kcl.ac.uk/policy-institute/assets/emerging-tensions.pdf>
23. Korinek, J., Moisé, E., Tange, J.: Trade and gender: A framework of analysis (OECD Trade Policy Papers, No. 246). OECD Publishing (2021). <https://doi.org/10.1787/6db59d80-en>
24. Lagesen, V.A., Pettersen, I., Berg, L.: Inclusion of women to ICT engineering – Lessons learned. *Eur. J. Eng. Educ.* **47**(3), 467–482 (2022). <https://doi.org/10.1080/03043797.2021.1983774>
25. Lim, L., Wang, Y.: Closing the gender divide in tech: Challenges and achievements in vogue. In: Marcus, A. (ed.) *Design, user experience, and usability: Users and interactions (DUXU 2015)*. Lecture Notes in Computer Science (Vol. 9187), pp. 228–238. Springer (2015). https://doi.org/10.1007/978-3-319-20898-5_22
26. Maltese, A.V., Tai, R.H.: Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among U.S. students. *Science Education* **95**(5), 877–907 (2011). <https://doi.org/10.1002/sce.20441>

27. McKinnon, M.: The absence of evidence of the effectiveness of Australian gender equity in STEM initiatives. *Aust. J. Soc. Issues* **57**(1), 202–214 (2020). <https://doi.org/10.1002/ajs.4.142>
28. Meier, P., Celis, K.: Sowing the seeds of its own failure: Implementing the concept of gender mainstreaming. *Soc. Polit. Int. Stud. Gend. State Soc.* **18**(4), 469–489 (2011). <https://doi.org/10.1093/sp/jxr020>
29. Monteiro, R., Lopes, M.: Closing the digital gender gap in Portugal: Challenges and achievements. In: *Algoritmos, teletrabajo y otros grandes temas del feminismo digital*. Dykinson (2023)
30. Monteiro, R., Lopes, M.A., Vieira, C.C., Santos, C.C., Ferreira, V.: Políticas públicas de igualdade de género na educação em Portugal: Da legislação à difícil transposição prática. Centro de Estudos Sociais, Universidade de Coimbra (2024). <https://estudogeral.uc.pt/handle/10316/116779>
31. OECD: OECD Digital Economy Outlook 2024 (Volume 1): Embracing the technology frontier. OECD Publishing (2024a). <https://doi.org/10.1787/a1689dc5-en>
32. OECD: OECD Digital Economy Outlook 2024 (Volume 2): Strengthening connectivity, innovation and trust. OECD Publishing (2024b). <https://doi.org/10.1787/3adf705b-en>
33. OECD: PISA 2022 database (2024c). <https://www.oecd.org/pisa/data/2022database>
34. Paterson, S.: What's the problem with gender-based analysis? Gender mainstreaming policy and practice in Canada. *Can. Public Adm.* **53**(3), 395–416 (2010). <https://doi.org/10.1111/j.1754-7121.2010.00134.x>
35. Prešički, T.: Silencing women in the digital age (2020). https://www.researchgate.net/publication/337776308_Silencing_women_in_the_digital_age
36. Radaelli, C.M.: Europeanisation: Solution or problem? *European Integration Online Papers* **8**(16) (2004). <https://doi.org/10.2139/ssrn.602662>
37. Rees, T.: Pushing the gender equality agenda forward in the European Union. In: Sagaria, M.D. (ed.) *Women, universities, and change: Gender equality in the European Union and the United States*, pp. 7–23. Palgrave Macmillan (2007). https://doi.org/10.1057/9780230603509_2
38. República Portuguesa: Estratégia Digital Nacional. XXIV Governo Constitucional (2024). <https://bo.digital.gov.pt/api/assets/etic/c508799d-8731-4a1f-b828-daac504c87e1/>
39. Simões, I.M.: Para um ensino das STEM e das TIC com perspectiva de género: Uma análise de práticas e de ausências (2024). <https://estudogeral.uc.pt/handle/10316/117095>
40. Stout, J.G., Dasgupta, N., Hunsinger, M., McManus, M.A.: STEMing the tide: Using ingroup experts to inoculate women's self-concept in STEM. *J. Pers. Soc. Psychol.* **100**(2), 255–270 (2011). <https://doi.org/10.1037/a0021385>
41. Su, R., Rounds, J.: All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields. *J. Pers. Soc. Psychol.* **109**(4), 543–566 (2015). <https://doi.org/10.1037/pspp0000027>
42. The Economist: Why young men and women are drifting apart. *The Economist* (2024). <https://www.economist.com/international/2024/03/13/why-the-growing-gulf-between-young-men-and-women>



Integrating Ethics and Gender Equality in Artificial Intelligence Education: A Study of Higher Education in Portugal

Camila Marques¹ , Rosa Monteiro¹ , and Nuno Lourenço² 

¹ Faculty of Economics of the University of Coimbra, Av. Dias da Silva, 165, 3004-512 Coimbra, Portugal

uc2020227742@student.uc.pt, rosa.monteiro@fe.uc.pt

² CISUC/LASI, DEI, University of Coimbra, R. Miguel Bombarda, 3030-790 Coimbra, Portugal
naml@dei.uc.pt

Abstract. Artificial Intelligence (AI) systems have the potential to replicate and even intensify existing social inequalities, particularly those related to gender. As AI technologies become ubiquitous, it is essential that developers are equipped to consider their ethical and societal implications. This study examines how ethics and gender equality are addressed in AI education within higher education institutions in Portugal. Through a qualitative, multi-phase methodology—including a desk-based review, an online questionnaire, and interviews—it investigates whether, where, and how these themes are integrated into AI-related university curricula, and how academic programs prepare future professionals to critically engage with these crucial issues. It analyzes the pedagogical approaches employed and the contents addressed, while also identifying institutional initiatives aimed at promoting ethical and gender-sensitive AI education.

Preliminary findings reveal that only 24% of AI-related programs explicitly incorporate a gender perspective, often indirectly through broader concepts such as fairness, bias, or justice. Explicit references to gender are scarce and typically limited to recommended bibliographies. Furthermore, these topics appear more prominently at the undergraduate level, with a marked absence in the doctoral programs analyzed.

These findings highlight critical gaps in AI education and the pressing need for more systematic integration of ethics and gender perspectives. By analysing how these topics are addressed in AI education in Portugal, this study contributes to the ongoing dialogue about ethical AI development. It calls for a comprehensive, forward-thinking approach that equips students not only with technical skills but also with the capacity to navigate the broader societal impact of the technologies they create.

Keywords: Higher Education · Gender · Ethics · Artificial Intelligence

1 Introduction

Artificial Intelligence (AI) is changing the ways we live, learn, work, and connect with one another. As intelligent systems increasingly permeate critical sectors, including healthcare, law, education, and security, it becomes evident that the design, development, and application of these technologies are not neutral processes. Rather, they reflect the values, biases, and limitations present in the data, design processes, and even developers. [18, 21, 22]. Contemporary AI encompasses a wide range of technologies, including machine learning, deep learning, neural networks, and natural language processing, with generative models like GPT representing the forefront of recent advances. While these systems vary in technical architectures, scholars such as Crawford [9] and UNESCO and IRCAI [31] emphasize that they commonly embed social biases, particularly regarding gender, due to skewed training data and lack of diversity in development teams. In this context, the integration of ethics and gender equality in AI production and analysis emerges as essential dimensions to ensure that AI technologies serve the interests of society as a whole, rather than perpetuating or exacerbating existing inequalities [1, 15, 23].

The incorporation of ethical principles into technological advancement seeks to promote transparency, justice, accountability, and inclusivity. Moreover, embedding a gender perspective within AI development enables the identification and rectification of issues such as algorithmic discrimination, the marginalization of diverse experiences, and the reinforcement of stereotypes within datasets and algorithms [1, 21]. These considerations are critical in fostering responsible AI practices that respect and uphold human rights.

Despite the growing recognition of the importance of these ethical and gender-related concerns, their integration into AI curricula in higher education institutions remains insufficiently addressed. Often relegated to secondary status or entirely absent from academic programs, the ethical and social dimensions of AI are frequently overshadowed by purely technical training or abstract philosophical ethics [6, 14, 19].

This trend is particularly troubling given the crucial role of higher education in fostering the critical competencies required for responsible AI technologies. To prevent the exacerbation of existing gender inequalities in the field, it is vital that AI education addresses issues of equality and non-discrimination [3, 14, 19]. Considering that AI remains a highly masculinized area and that discriminatory biases are often embedded in AI algorithms, proactive measures are needed to prevent these outcomes [14]. In this context, there is a growing consensus on the importance of incorporating ethics and, in particular, a gender perspective, into AI curricula, as a condition to prepare future professionals capable of creating ethical and socially responsible AI systems [3, 14, 19].

This study focuses on Portugal, where systematic research on the incorporation of ethics and gender equality within AI-related higher education degrees is notably absent, highlighting a significant gap in knowledge and action. By examining this specific context, the study aims to contribute to a broader critical dialogue involving higher education institutions, educators, students, and policymakers.

The primary objective of this article is to analyze the extent to which ethical and gender equality issues are addressed in AI-related university degrees across Portugal.

This analysis will identify when these issues are introduced throughout the academic journey, assess whether they are addressed comprehensively or in isolation, and examine the teaching methodologies and specific content through which they are conveyed. Simultaneously, this study will map relevant initiatives undertaken by institutions and educators that foster a more ethical, inclusive, and socially responsible approach in higher education institutions related to AI teaching and research.

The organization of this article is as follows: the next section presents the theoretical framework, reviewing existing literature on the integration of ethics and gender equality in AI education. This is followed by a description of the adopted methodology of the global study, which encompasses a desk survey, an online questionnaire, and interviews. The fourth section details the main findings, from the data gathered from the desk survey and the questionnaire, while section five offers concluding remarks on AI education in Portugal.

2 Literature Review

AI systems have been widely criticized for perpetuating and amplifying gender inequalities across multiple domains. One clear manifestation is the reinforcement of gender stereotypes, exemplified by the design of female-voiced virtual assistants such as Siri and Alexa, which reproduce traditional gender roles by portraying women as servile, obedient, and constantly available [25]. This design choice not only reinforces patriarchal norms but also erases alternative representations of women in the digital sphere. Algorithmic discrimination further exacerbates inequalities, as shown by cases like Amazon's recruitment system, where female candidates were systematically downgraded due to biased historical training data [11]. In the healthcare sector, AI models often rely on male-centered datasets, leading to critical oversights in medical diagnostics for women. The UNESCO and IRCAI [31] report underscores how large language models (LLMs) reinforce these patterns, systematically associating women with domestic roles while assigning men to high-status professions. A persistent lack of diversity in AI development teams, particularly the underrepresentation of racialized women, limits the inclusiveness of technological outcomes, reinforcing what has been termed the "algorithmic glass ceiling". Facial recognition tools also present higher error rates for women, especially women of color, increasing their exposure to surveillance and criminalization [5]. Moreover, the automation of labor threatens female-dominated sectors, accentuating economic disparities [27]. Together, these dynamics illustrate how AI, far from being a neutral tool, is a powerful agent in reproducing and deepening existing gendered power structures [18, 21]. This literature review will explore three key topics: first, international regulations and recommendations on ethics and gender in AI, focusing on frameworks from organizations such as UNESCO and the European Union. Next, we will examine algorithmic discrimination, which highlights how biases in data and design can lead to unfair outcomes. Finally, we will discuss the importance of gender pedagogy in AI education, advocating for gender-sensitive approaches in teaching the next generation of AI professionals.

2.1 International Regulations and Recommendations on Ethics, Gender and Artificial Intelligence

This section explores international regulations and recommendations regarding ethics, gender, and AI. Organizations like UNESCO and the European Union have put forward frameworks that emphasize the importance of considering ethics and a gender perspective in AI development [17, 28, 29].

At the global level, UNESCO's Artificial Intelligence and Gender Equality report [28] proposes that gender equality should be recognized as an autonomous principle in AI, recommending a comprehensive integration of the gender perspective throughout the entire AI lifecycle. It also emphasizes the importance of education and training, particularly through the development of curricula that bridge social sciences, ethics, and technological literacy. This approach is reinforced by the Recommendation on the Ethics of Artificial Intelligence [29], which urges member states to secure funding for gender-sensitive AI systems, promote female participation at all stages of development, and incorporate gender considerations into ethical impact assessments. In the field of education, it further recommends the development of AI ethics curricula across all levels of education and the training of researchers to embed ethical concerns into their projects.

At the European level, the Ethics Guidelines for Trustworthy AI, developed by the EU High-Level Expert Group on AI [17], outline seven key requirements, including diversity, non-discrimination, and fairness. These principles aim to ensure that AI systems respect human dignity and promote equality. Nevertheless, when drafting the AI Act [12], many of these concerns were not adequately addressed [10]. In fact, the final version of the AI Act, which classifies AI systems based on their level of risk (minimal, limited, high, and unacceptable), faced criticism for lacking explicit references to gender equality and non-discrimination, elements that were included in earlier drafts, and for offering limited protection to vulnerable groups [10].

In addition to the AI Act, the General Data Protection Regulation (GDPR) [13] serves as a foundational regulation for all technology, emphasizing principles such as lawfulness, data minimization, accuracy, and accountability. However, similar to the AI Act, it also fails to specifically address the gender dimension.

In Portugal, Law no. 58/2019 of 8 August ensures the implementation of the European General Data Protection Regulation (GDPR) within national law. The National Data Protection Commission (CNPd) was appointed as the national supervisory authority for the purposes of the GDPR and this law, as well as for other legal and regulatory provisions concerning the protection of personal data [2].

On a more positive note, the Council of Europe's Framework Convention on AI and Human Rights [8] fortifies the discussion around gender equality by introducing guiding principles that mandate the guarantee of gender equality in AI systems, emphasizing human dignity, equality, and non-discrimination.

In conclusion, while international policy frameworks increasingly recognize the importance of gender equality in AI, it is crucial to assess the extent to which these regulations and recommendations are being implemented in practice, particularly in addressing the gender inequalities previously discussed.

2.2 Algorithmic Discrimination

Discrimination in AI systems is a well-documented issue, appearing in various forms across everyday products and applications. One of the most prominent expressions of gender inequality in AI is algorithmic bias, which is both produced and reinforced by these systems. This bias can take multiple forms, including discrimination stemming from biased training data, proxy variables, disparate impact, and targeted advertising [25]. Other notable examples include facial recognition software with high error in racialized and women's faces [5] and feminized virtual assistants that reinforce gender stereotypes [30].

These biases and discrimination come from factors like historical biases in data, lack of diversity in the development teams, and the opacity of algorithmic systems [20, 21]. Analysis emphasized the importance of considering the social and historical contexts in which algorithms are designed and deployed, as well as the role of human choices and values in shaping algorithmic outcomes. Noble [25] uses the term “algorithmic oppression” to reinforce the gender and intersectional inequality complex impacts of AI.

The origin of these biases goes beyond the simple demographic composition of technical teams. While the under-representation of women and racial/ethnic minorities in technological production is a relevant factor [1], studies show that technical processes themselves, such as data labelling, design decisions and evaluation metrics, are socially constructed and carry normative assumptions [18, 21].

Another example of discrimination within the field of AI is the systematic exclusion of invisible workers, mostly women, involved in data preparation exemplifies the phenomenon of *fauxtimation*, coined by Astra Taylor [27], and describes the way automation hides essential human labor.

Various strategies have been proposed to combat algorithmic discrimination. Literature reveals that computational solutions and techniques prevail [3]. They consist of pre-processing, processing and post-processing approaches that seek to mitigate bias in data and AI systems stand out [3, 22, 32]. Tools such as AI Fairness 360 and Fairlearn enable the measurement and mitigation of algorithmic bias at various stages of the machine learning lifecycle. However, these technical solutions face significant limitations. They often ignore the structural causes of discrimination, they can reinforce inequalities by searching for more diverse datasets in an ethically questionable way, and they present a trade-off between accuracy and fairness [4, 18, 25, 32].

Critical literature thus proposes more comprehensive approaches, such as participatory design and value-sensitive design [3], which implies a comprehension of the mechanisms of inequality present in the data itself, for instance. Authors such as Safiya Noble [25] advocate the need for an “algorithmic redemption” that goes beyond computational equity and directly confronts structures of domination and oppression. A socio-technical, multidisciplinary and feminist approach is thus crucial to understanding and correcting the ethical and gender impacts of AI in real contexts [20, 21].

2.3 Gender Pedagogy in Artificial Intelligence Teaching

The integration of a gender perspective in AI teaching has been receiving little interest. Studies in Spain reveal that the ethical and social aspects of AI are introduced as a

transversal competence in a very limited number of curricular units, or only in a specific unit focused on legal aspects [6]. Regarding the gender perspective, it is included in only one of the four undergraduate degrees analyzed, as a transversal competence [6]. As for master's programs, transversal competences related to ethics and social commitment in AI are either not introduced at all or included in a very limited way. Moreover, none of the master's programs explicitly incorporate a gender perspective as a transversal competence [14].

Gender sensitive pedagogies and the integration of a gender mainstreaming approach in education has been recommended in international normatives for more than three decades now [24]. Resulting from feminist pedagogies, gender sensitive pedagogies, aims to an antisexist, non-hierarchical teaching which highlights women experiences, and includes awareness and empowering strategies to identify and resist oppression structures, emphasizing experience-based learning models [24]. In the domain of AI, recommended pedagogical practices include the integration of specific ethics curricular units with a gender focus, classroom discussion on topics such as the ethics of AI systems and the dangers associated with AI development [7], the use of real-life scenarios that address data validity and social inequalities [19]. It is also proposed to introduce Science, Technology and Society (STS) knowledge into curricula and to adopt an integrated institutional approach (whole school approach) that involves the entire academic community in promoting gender equality [19]. These proposals were used as categories in our analysis of current pedagogical approaches used in Portuguese higher education institutions.

Recommended pedagogical practices point to the need for more systematic and interdisciplinary curricular approaches that include ethics, gender studies and the active participation of the entire academic community.

3 Methodology

This article is part of a major study which aims to analyze the integration of ethical issues and gender equality in the teaching of AI within higher education institutions in Portugal. To achieve this overarching goal, the study outlined five specific objectives: (1) to identify and examine the curricular units of higher education degrees focused on AI and Data Science (DS) that incorporate ethical and social issues related to AI; (2) to investigate the extent in which gender perspectives are integrated into these curricular units; (3) to analyze the perceptions of lecturers and degree coordinators regarding the significance of including ethical issues and gender perspectives into AI education, as well as the pedagogical methods employed; (4) to assess existing institutional initiatives, policies and practices within Portuguese higher education institutions aimed at promoting gender equality and enhancing the female participation in the field of AI; and (5) To identify perceived challenges and opportunities associated with incorporating gender perspectives into AI education, drawing from the experiences of professionals in the field.

The study employs a mixed-methods approach, combining qualitative and quantitative methods to explore how ethical issues and gender equality are integrated into AI education. The methodological framework is structured across three complementary phases and tools: desk survey, online questionnaire and interviews.

The first consisted of a desk survey, with the goal of analysing the presence of ethics and gender equality perspectives in AI-related university degrees and curricular units offered by higher education institutions in Portugal. The primary data source was the portal of the Directorate-General for Higher Education (DGES), utilizing filters based on the National Classification of Education and Training Areas (CNAEF). Only courses specifically dedicated to Artificial Intelligence (AI) and Data Science (DS) were included, covering all three academic cycles. The inclusion of DS courses is justified by their close relationship to AI, making their analysis equally relevant.

In total, 37 courses were identified: 14 from the first-cycle (bachelor's) programs, 22 from the second-cycle (master's) programs, and 1 from the third-cycle (doctoral) programs. For each course, the curricular units and respective syllabuses (2024/2025 academic year, or the most recent available data) related to ethics, social impact and gender were gathered from institutional websites and analyzed in detail. Considering that explicit references to the gender perspective are still very limited, a list of keywords was used to detect possible implicit mentions: equity, equality, justice, inclusion, non-discrimination, non-prejudice, diversity, representativeness, empowerment and shared benefits [16].

To complement the information available on the institutional websites, an online survey was designed to explore how gender-related issues are approached in AI education, particularly in courses related to AI ethics and responsibility. The goal was to understand whether topics such as gender inequality and algorithmic bias are included in the curriculum, the teaching methods employed, and the relevance attributed to these topics by course coordinators and AI professors.

The survey targeted 40 coordinators and lecturers at university degrees in AI and DS, including those teaching subjects related to ethics or responsibility in AI. It served as a complementary source of information to enrich the data collected from the websites consulted. The questionnaire received 13 responses: ten from coordinators of AI and DS programs, one from a lecturer in ethics and social issues in AI, and two from AI lecturers.

In addition to the documentary analysis and survey, qualitative interviews were conducted with key experts in AI and ethics, including professors and researchers. These interviews provided further insights into the current debate on integrating gender perspectives into AI curriculums, effective pedagogical methods, and the structural challenges that impede such integration. Additionally, the interviews examine the role of interdisciplinary collaboration and the influence of national and international ethical guidelines, while also identifying institutional initiatives and proposing concrete measures to foster a more inclusive and socially responsible approach to AI education.

In this paper we only present preliminary analysis of the data gathered from the desk survey and the additional data collected through the survey with coordinators and lecturers.

4 Results

This section presents the preliminary analysis of the data collected, structured into two main areas. The first focuses on institutional commitments and actions related to gender equality and ethics. The second explores how ethical and social dimensions, including

the gender perspective, are addressed within AI-related curricula. Together, these areas offer an initial overview of how institutions and academic programs engage with the ethical and gender-related issues in AI education.

4.1 Institutional Efforts Toward Gender Equality and Ethics

The data collected from 37 AI-related programs reveal a significant gender imbalance in both the coordination and composition of teaching staff within university degrees focused on AI. Notably, 73% of these programs are coordinated by men. Among the 19 degrees for which detailed data were available, men make up 66% of the lecturers, while women account for just 34%.

In addition, there is a marked lack of training on gender equality for teaching staff, particularly regarding its relevance and application in the context of AI. Faculty engagement in initiatives related to responsible AI, diversity, and gender equality also appears limited: only eight program coordinators reported faculty involvement in such efforts, mainly linked to external partnerships.

Additionally, discussion surrounding gender equality and diversity in AI is infrequent among teaching teams, during meetings or supervision of dissertations. Although there are some specific initiatives to promote gender equality in the areas of technology and engineering, there are still few structured actions promoted by higher education institutions, particularly by computer science and engineering departments, to integrate the gender perspective into AI.

In summary, while there are few examples of good practices and initiatives, the majority of higher education institutions in Portugal still lack coordinated and structured efforts to promote inclusion and gender equality as a pedagogical project in the teaching of AI. There is a disconnection between efforts aimed at promoting equality, often implemented through external activities or projects, and the content and methodologies employed in the classroom. As a result, the presumed “neutrality” of teaching practices inadvertently reinforces the very inequalities that these institutions seek to challenge through their external commitments.

4.2 Curricular Integration of Ethical and Social Issues in AI Education

To further investigate the integration of ethical and social issues in AI education in Portugal, we conducted a more in-depth analysis of curricular units that incorporate these dimensions. In total, 39 curricular units were analyzed, including 27 specifically related to ethical and social issues, and 12 that incorporate these issues along with other contents.

From the 37 AI-related programs identified, it was found that 70% (26) include at least one curricular unit specifically dedicated to ethical and social issues. However, this inclusion varies significantly across academic cycles: 93% (13) of undergraduate programs include such a unit, compared to only 55% (12) of master's programs, while the only identified doctoral program also includes at least one.

Despite the apparent comprehensiveness of some programs, only 38% of the degrees that include a specific curricular unit on the ethical and social dimensions of AI go beyond

a narrow focus on data security and (legal) protection. We identified 16 curricular units, out of the 27 specifically dedicated to ethical and social impacts of AI, that address topics such as the risks and benefits of AI (75%), ethical foundations (69%), and responsible AI (63%). Even though these units provide a more comprehensive exploration of the problem, some aspects are still addressed in a limited way. Additionally, the most common pedagogical approaches employed in these 16 curricular units include lectures (75%) and case study analyses (50%). Most of these curricular units are also compulsory (81%) and boast a relatively balanced gender distribution among the teaching staff (56% men and 44% women). In terms of scientific areas, these curricular units are predominantly situated in technological disciplines (44%), with the remaining units associated with Social Sciences, Law and Philosophy.

A comparison between academic cycles reveals notable disparities in their focus and content. The undergraduate degrees are distinguished by their broader diversity across various scientific disciplines, while the master's and doctoral programs predominantly concentrate on technological areas. Regarding content, the first and second cycles offer broader coverage of ethical and social issues, with undergraduate programs emerging as the most comprehensive overall.

In addition to the 16 curricular units specifically dedicated to the ethical and social implications of AI, we identified 12 units that address these issues as part of a broader curriculum, rather than as their central focus. These units typically integrate ethical and social considerations alongside technical or methodological content, and are particularly concentrated at the master's level, such as Seminars and Specialization Ateliers. They tend to involve a higher number of lecturers from the Social Sciences and Humanities and employ slightly more diverse pedagogical approaches.

In short, despite the presence of at least one curricular unit related to ethical and social considerations in the majority (70%) of the 37 AI-related university degrees, only 38% (14) delve into topics beyond privacy and data protection. Much of the discourse centers on vital issues such as security, data protection, and the risks and benefits associated with AI. However, critical topics like prejudice and algorithmic discrimination are still significantly underrepresented.

The analysis focused on the presence of a gender perspective in the syllabus of the curricular units that we have just identified, it was found that only 24% (9) of the university degrees incorporate this perspective in their teaching, with a more significant presence at the undergraduate programs, where 36% (5) address gender issues, compared to only 18% (4) of master's degrees, and an absence in the sole doctoral program analyzed. Additionally, two degree programs, one undergraduate and one master's, were found to explicitly address gender issues.

In total, ten curricular units that address gender issues were identified. In most instances, the gender perspective is incorporated through dedicated units that provide a more comprehensive examination of the ethical and social dimensions of AI (8). Only two degree programs integrate this perspective within broader curricular units that also cover other themes.

In terms of their distribution across academic years, 80% (4) of the curricular units in undergraduate degrees that integrate the gender perspective are concentrated in the

third year. In master's programs, they predominantly occur in the first year, accounting for 60% (4) of the cases.

The teaching staff stands out not only for the diversity of their academic and professional backgrounds, but also for their active involvement in prominent scientific and technological initiatives, such as the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM), and they contribute to interdisciplinary projects like the Artificial Intelligence, Data and Algorithms Social Sciences Research Network (AIDA). In addition to their academic and research commitments, they are also engaged in efforts to promote inclusion and diversity in technology through external initiatives such as “As Raparigas do Código” and the advancement of responsible AI, exemplified by projects like “NextGenAI – Centre for Responsible AI”.

The analysis also indicates that most curricular units make implicit references to gender issues. While gender concerns are addressed, they are not explicitly stated. Instead, the language frequently used includes terms like “fairness”, “bias”, “(un)equality”, “(non-)discrimination”, “equity”, and “diversity”.

Two exceptions stand out: the curricular units “Technology, Economy and Society” at Iscte - University Institute of Lisbon (undergraduate) and “Artificial Intelligence and Society” at the University of Porto (master's). In the former, explicit references to gender are found in the recommended bibliography, referencing an article addressing sexual harassment in virtual reality and the Proteus effect, which examines how digital representations influence individuals [26]. Similarly, the latter includes an explicit mention of gender in the complementary bibliography, notably in the work “Data Feminism” by Catherine D'Ignazio and Lauren F. Klein [11], which explores gender and feminism in the field of computational design.

In addition, it is possible to identify that references to gender appear mainly in the course contents and are rarely included in the teaching objectives or methods. After the contents, the recommended bibliography is the part of the syllabus with the most references to gender, as well as being the only place to explicitly include the topic.

5 Final Remarks

In summary, the analysis demonstrates that the integration of a gender perspective in curricular units addressing ethical and social issues in AI remains limited and predominantly implicit, in Portugal. Only 24% (9) of AI-related university degrees include this perspective, which is more common at the undergraduate level and entirely absent in the examined doctoral program. The approach to gender is largely mediated through broader concepts such as justice, bias, and equity, with explicit references being scarce and primarily located in recommended bibliographies.

This trend indicates that while there are efforts to incorporate gender issues into AI, there is still a considerable distance to cover before this perspective can be integrated more directly and systematically into academic curricula.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Armutat, S., et al.: Artificial Intelligence – Gender-Specific Differences in Perception, Understanding, and Training Interest. *icgr*. **7**(1), 36–43 (2024). <https://doi.org/10.34190/icgr.7.1.2163>
2. Assembleia da República: Lei n.º 58/2019, de 8 de agosto. (2019)
3. Barnes, E., Hutson, J.: Navigating the ethical terrain of AI in higher education: Strategies for mitigating bias and promoting fairness. *Forum Edu. Stud.* **2**(2), 1229 (2024). <https://doi.org/10.59400/fes.v2i2.1229>
4. Benjamin, R.: *Race after technology: abolitionist tools for the new Jim code*. Polity, Cambridge, UK Medford, MA (2020)
5. Buolamwini, J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Presented at the 1st Conference on Fairness, Accountability and Transparency (2018)
6. Cernadas, E., Calvo-Iglesias, E.: Gender perspective in Artificial Intelligence (AI). In: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality, pp. 173–176. ACM, Salamanca Spain (2020). <https://doi.org/10.1145/3434780.3436658>
7. Cernadas, E., Fernandez-Delgado, M.: Embedded ethics to teach machine learning courses: an experience. In: 2021 XI International Conference on Virtual Campus (JICV), pp. 1–4. IEEE, Salamanca, Spain (2021). <https://doi.org/10.1109/JICV53222.2021.9600426>
8. Council of Europe: Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (2024)
9. Crawford, K.: *Atlas of AI: power, politics, and the planetary costs of artificial intelligence*. Yale University Press, New Haven London (2021)
10. Custers, B.H.M., Fosch-Villaronga, E. (eds.): *Law and artificial intelligence: regulating AI and applying AI in legal practice*. Asser Press, The Hague (2022)
11. D'Ignazio, C., Klein, L.F.: *Data Feminism*. The MIT Press, Cambridge (2020)
12. European Union: AI Act. (2024)
13. European Union: Regulamento Geral sobre a Proteção de Dados. (2016)
14. García, E.C., Calvo, E.: Perspectiva de género en Inteligencia Artificial, una necesidad. *Cuestiones de género* **17**, 111–127 (2022). <https://doi.org/10.18002/cg.i17.7200>
15. González, A.S., Rampino, L.: A design perspective on how to tackle gender biases when developing AI-driven systems. *AI Ethics*. **5**(1), 201–218 (2025). <https://doi.org/10.1007/s43681-023-00386-2>
16. Guevara-Gómez, A., et al.: Feminist perspectives to artificial intelligence: Comparing the policy frames of the European Union and Spain. *IP*. **26**(2), 173–192 (2021). <https://doi.org/10.3233/IP-200299>
17. HLEG: Orientações éticas para uma IA de confiança. Publications Office, LU (2019)
18. Hoffman, S.G., et al.: Five Big Ideas About AI. *Contexts* **21**(3), 8–15 (2022). <https://doi.org/10.1177/15365042221114975>
19. Joyce, K. et al.: Engaging STEM Ethics Education. *Engaging STS*. **4**, 1–7 (2018). <https://doi.org/10.17351/ests2018.221>
20. Joyce, K., Cruz, T.M.: A Sociology of Artificial Intelligence: Inequalities, Power, and Data Justice. *Socius: Sociological Research for a Dynamic World*. **10**, 23780231241275393 (2024). <https://doi.org/10.1177/23780231241275393>
21. Klein, L., D'Ignazio, C.: Data Feminism for AI. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 100–112 ACM, Rio de Janeiro Brazil (2024). <https://doi.org/10.1145/3630106.3658543>

22. Leavy, S. et al.: Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 695–703. ACM, Virtual Event USA (2021). <https://doi.org/10.1145/3461702.3462598>
23. Leavy, S.: Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. Presented at the ACM/IEEE 1st International Workshop on Gender Equality in Software Engineering. New York, NY (2018). <https://doi.org/10.1145/3195570>
24. Monteiro, R.F.B.L.: Resistências institucionais e antigênero à igualdade de gênero na educação em Portugal. *Cad. Pesqui.* **54**, e11410 (2024). <https://doi.org/10.1590/1980531411410>
25. Noble, S.U.: Algorithms of oppression: how search engines reinforce racism. New York University Press, New York (2018)
26. Patel, N.J.: Reality or Fiction - Sexual Harassment in VR, The Proteus Effect and the phenomenology of Darth Vader — and other stories, <https://medium.com/kabuni/fiction-vs-non-fiction-98aa0098f3b0>. Last accessed 03 March 2025
27. Taylor, A.: The Automation Charade, <https://logicmag.io/failure/the-automation-charade/> (2018)
28. UNESCO: Artificial intelligence and gender equality: Key findings of UNESCO’s global dialogue (2020)
29. UNESCO: Recomendação sobre a Ética da Inteligência Artificial (2022)
30. UNESCO, EQUALS Skills Coalition: I’d blush if I could: closing gender divides in digital skills through education. UNESCO (2019). <https://doi.org/10.54675/RAPC9356>
31. UNESCO, IRCAI: Challenging systematic prejudices: an investigation into bias against women and girls in large language models (2024)
32. Valentim, I., et al.: The Impact of Data Preparation on the Fairness of Software Systems. In: 2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE), pp. 391–401. IEEE, Berlin, Germany (2019). <https://doi.org/10.1109/ISSRE.2019.00046>



Gender and Emerging Digital Technologies in Education

Davorka Radaković¹(✉)  and William Steingartner²(✉) 

¹ Faculty of Sciences, Department of Mathematics and Informatics,
Division of Informatics, University of Novi Sad, Novi Sad, Serbia
`davorkar@dmf.uns.ac.rs`

² Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Košice, Slovakia
`william.steingartner@tuke.sk`

Abstract. The integration of digital technologies, content and processes is reshaping all sectors of education: from curriculum reform to teacher assessment and development. Among these technologies, applications of artificial intelligence (AI) present significant opportunities to improve the learning experience, especially in secondary and higher education. In this study, we focus on the potential impact of AI on teaching and learning by examining how students perceive AI as a learning tool. We also examine how AI applications are integrated into academic processes and activities. Building on previous research and in line with the European Education Area (EEA) initiative, the study uses a cross-sectional analysis to assess the benefits and challenges of AI in educational contexts. The findings provide insights into the effectiveness of AI in supporting communication, programming, content creation and explaining concepts, while highlighting the importance of developing skills for educators and students to navigate the evolving AI landscape.

Keywords: AI · ChatGPT · Education · Teaching and Learning · Student Engagement · Ethics

1 Introduction

Digital technologies, content and processes influence all education sectors: schools, higher education, and informal learning; shaping the entire value chain, from curriculum reform to teaching, assessment, and teacher development, while engaging educators, learners, and school leaders.

Applications based on Artificial Intelligence (AI) have a huge contribution in everybody's daily life. In addition, AI offers a wide range of potential benefits for students and teachers. AI in education poses challenges, demanding research and practical application. Effective data use, ethics, and AI advancement require skill development for teachers, administrators, and students.

ChatGPT can assist students in: developing communication skills in a foreign language [6, 9]; in programming [21]; creating content, answering questions, explaining concepts, and giving suggestions [10, 25].

Since the education sector has benefited enormously through the integration of tools and platforms driven by digital technology, many authors analyze modern AI tools available to learners and educators [14, 16, 20].

The European Education Area (EEA) [1] initiative builds more resilient and inclusive national education systems that involve them with digital learning and pedagogies, and enables policy makers to design, implement, and evaluate programs and projects for the integration of digital learning technologies in educational and teaching systems.

This study is inspired by research conducted by [5]. The purpose of this research is to determine how AI can impact on teaching and learning in middle and higher education. The study aimed to examine students' attitudes toward AI through a cross-sectional analysis. Our research questions centered on two key aspects of AI use, particularly in relation to ChatGPT. First, we investigated whether and to what extent students perceive AI as a useful tool for their learning. Second, we examined how students integrate AI into their academic activities and the potential challenges they encounter in doing so.

The rest of the paper is organized as follows. Section 2 reviews related work. Further, we present the review method following the results of this review with a discussion of our results. Finally, in Sect. 5 we conclude the paper with a brief summary and discuss the future work.

2 Related Work

In this paper we focus on examining, analyzing, and drawing useful conclusions about using AI tools in educational environment in higher education. Our literature review shows that researchers have extensively examined the potential of AI's transformative role in teaching and learning in various subjects.

The EU's focus on digital transformation and the use of AI in education, as evidenced by the AI Act and the Coordinated Plan on AI. One of the European Commission framework projects is SELFIE [3] (Self-reflection on Effective Learning by Fostering the use of Innovative Educational technologies). It is a free tool designed to help schools embed digital technologies into teaching, learning, and assessment. Students, teachers, and school leaders to take part in SELFIE on an anonymous basis; and use the findings from SELFIE to see how technologies could support teaching, learning and student assessment. Digital learning technologies, in the context of DigCompOrg [2], constitute a key enabler for educational organizations, which can support their efforts to achieve their particular mission and vision for quality education.

In article [22], La Roca et al. illustrate the integration of a chatbot into a first-year course in the Computer Engineering. Students emphasized that a key advantage was the chatbot's 24/7 availability and its intuitive, user-friendly interface.

Qualitative research [11] provides an overview of the effectiveness of using ChatGPT as a virtual assistant. It helps improve student productivity in higher education by providing useful information and resources, helping to improve

language skills, facilitating collaboration, improving time efficiency and effectiveness, and providing support and motivation. Baidoo-Anu and Ansah have discussed in [5] about the benefits of ChatGPT and related generative AI in advancing teaching and learning. Also, they have listed possible drawback such as generating wrong information, biases in data training, which may augment existing biases, privacy issues etc.

On the other hand, the study [17] showed that ChatGPT did not improve performance in various disciplines. Mike et al. investigate ChatGPT's impact on effectiveness, efficiency, and problem solving among higher education students in law, business informatics, and media and communication. Similarly, Wu in [26] expresses concern that students receiving very fast answers from ChatGPT have a wrong idea about how much they have mastered and learned, and this can lead to an ineffective learning process that exceed the cognitive load capacity.

There are no many papers that observe using ChatGPT for teaching and learning in data science education, so paper [27] incorporates ChatGPT into a data science course and collects valuable insights from students, as well as sharing feedback from the instructor. The findings highlight the uniqueness of data science education while revealing new opportunities and challenges in integrating ChatGPT into the curriculum. Jindal et al. suggest to educators, policymakers, and technology developers that they have to work collaboratively to establish frameworks that promote responsible and ethical use of ChatGPT in education.

In study [9], study examines how ChatGPT supports learning the four core language skills—listening, reading, speaking, and writing—within German language education. The chatbots enable students to practice the foreign language anytime and almost anywhere, allowing them to engage in language practice. It is shown that ChatGPT can assist in the development of students' vocabulary, reading comprehension, improve writing skills etc. Focus on ChatGPT is also given through analysis of the significance and utilization of artificial intelligence in learning systems in higher STEM education [13]. Furthermore, the study [23] found that students viewed ChatGPT as an innovative set of AI tools that can be used in classroom settings, including programming and computer languages, promoting sustainable adoption of AI in education. In addition, the study [19] showed that the integration of GPT and innovative learning methods significantly improved academic performance and motivation.

Chen et al. [8] systematically reviews AI applications in education, including personalized learning, intelligent tutoring systems, and automated assessment. It also addresses technical, ethical, and pedagogical challenges, emphasizing the need for balanced integration of AI with human-centered teaching approaches.

Another overview of AI's integration into educational settings is given in [18]. Ng et al. discuss various applications such as intelligent tutoring systems, personalized learning platforms, and automated assessment tools, highlighting their potential to improve educational outcomes.

Similarly, Lameris et al. [15] discuss the evolution of AI applications: from early computer-based tools to advanced systems such as intelligent tutoring, adaptive learning platforms, and humanoid robots, highlighting their impact on

administrative efficiency, personalized instruction, and improved learning outcomes.

Furthermore, the review [16] categorizes AI applications in education into proactive (e.g. admissions, course scheduling, content generation) and reactive (e.g. knowledge delivery, performance assessment, outcome prediction) phases. It examines research trends over two decades, highlighting how AI methods support various educational stakeholders and discussing the evolution of solution approaches in the field. In addition, in the study [7] 300 German-speaking adults were questioned to develop a measurement instrument Meta AI Literacy Scale (MAILS) for AI literacy with the facets Use & apply AI, Understand AI, Detect AI, and AI Ethics and the ability to Create AI as a separate construct, and AI Self-efficacy in learning and problem-solving and AI Self-management.

Besides that, Guo et al. in [12] examine teachers' acceptance of AI (TAAI) in educational settings. The instrument assesses five dimensions: perceived usefulness, perceived ease of use, behavioral intention, self-efficacy, and anxiety.

Our approach to developing a questionnaire to assess AI literacy had gone in the similar directions, i.e. to determine general perception, awareness, and usage of AI tools in teaching and learning along with ethical considerations. The methodology applied and discussion on obtained results, are presented in the rest of the paper.

3 Methodology

The research data in this study was collected through the primary data collection: questionnaire [24]. Data were collected online using Google Forms [4] from students during the beginning of the summer semester of 2024/2025. The study took a random sample of students with similar backgrounds: computer science, information technologies, and mathematics.

Data were therefore collected from student classes that were available to the researchers at the time from universities in Poland and Slovakia. The questionnaire was written in English. Before taking part in the study, participants were informed of its general purpose and provided their informed consent.

The first section of the questionnaire gathered students' general demographic information such as age, gender, and educational status. The second section assessed perceived general perception, awareness, and usage of AI in students' field, while the third section measured attitude toward usage of AI tools. The fourth section observe the level of satisfaction while using them in teaching and learning, ethics, and confidence. Final part examines the consideration of the AI use for students who still have not used AI.

3.1 Participants

The study results presented in the article are obtained from a data sample of a total of 84 students from two universities: Cardinal Stefan Wyszyński University

in Warsaw, Poland (36 students), and Technical University of Košice, Slovakia (48 students).

The highest level of education completed by 87% of the students is high school (pre-university entry, usually taken at age 18), while 13% have completed a bachelor’s degree.

Table 1. Distribution of students by gender

Gender	No. of students		Total
	Cardinal Stefan Wyszyński University	Technical University of Košice	
Male	24	39	63
Female	11	8	19
Prefer not to say	1	1	2
Total	36	48	84

19 students considered themselves female (22.62%), while 63 students identified as male (75%). Two students prefer not to say their gender (2.38%). There was some concern about the gender ratio; however, it is common in these types of the study. Table 1 presents the description of the participants by gender in this study (total and by Universities).

Table 2. Distribution of students by age

Age range	No. of students		Total
	Cardinal Stefan Wyszyński University	Technical University of Košice	
18-19	12	17	29
20-23	18	31	49
24-29	4	/	4
30-37	2	/	2

In the present sample, the average age was 20.77 years (SD = 2.796 years, ranging from 18 to 37 years). Table 2 presents the distribution of the students by age range in this study (total and by Universities).

4 students are major in mathematics, 29 in computer science, and 51 in information technologies. Distribution of students by major subject and gender is shown in Table 3.

Table 3. Distribution of students by major subject and gender

Gender	No. of students	No. of students	No. of students
	Mathematics	Computer Science	Information Technologies
Male	24	25	36
Female	2	3	14
Prefer not to say	/	1	1
Total	4	29	51

Overall, 72.6% of all students consider that they have access to AI tools in their educational environment, 20.2% are not sure, while 7.1% consider that they do not have access. Correspondingly, similar ratio is between genders as shown in Table 4.

Table 4. Frequency of access to AI tools in your educational environment (by gender)

Gender	No	Not sure	Yes
Male	5	13	45
Female	1	3	15
Prefer not to say	/	1	1
Total	6	17	61

4 Results

In this section we present the results of statistical analysis that was performed on the collected data.

The survey questionnaire was crafted to gather in-depth insights from participants, including questions that explored the perceived benefits of using AI tools in learning and any challenges faced during their use.

4.1 General Perception, Awareness and Usage of AI Tools

The survey was designed to explore the attitudes of students to the use of AI tools in teaching and learning using the structure of multiple-choice questions including commonly applicable criteria, and at the same time inviting opinions through open-ended questions, and several Likert-type scales.

Students' answers to the question *How do you perceive the role of AI in your field?* are: positive - *Very helpful; Faster documentation, writing boiler plate code, helping with creativity; Very important; It helps me understand something*

new; and with reserving judgment -I see AI as a potential “double edged sword”; and it comes down to how one uses it...; helps to understand the topics that the teacher explained poorly.

Students rate how familiar are they with AI on a 5-point Likert scale with the anchors strongly “not at all familiar”/“slightly familiar”/“somewhat familiar”/“moderately familiar”/“extremely familiar”. In our sample, 50 students claimed to be familiar with AI in “moderately familiar” or “extremely familiar” way, thereof 8 female students.

On the other hand 8 students (9.5%) claimed that they have not used AI tools in their work or daily life, of which 2 are female. These students were forwarded to the final part of questionnaire. 4 students claim that would consider using AI, one student maybe, and 3 would not consider using in the future AI. Withal, 3 students responded they would use AI in an educational context, three were unsure, and two responded they would not. For both of the last questions, one female student answered “maybe” and another answered “no.”

4.2 Using AI Tools

These questions were answered by 76 students who claimed that have used AI. A high overall percentage of students (98.7%) reported using AI Assistants (Chatbots), followed by 57.9% who use Grammar and Writing Improvement tools, and Search Engines 47.4%(see Fig. 1).

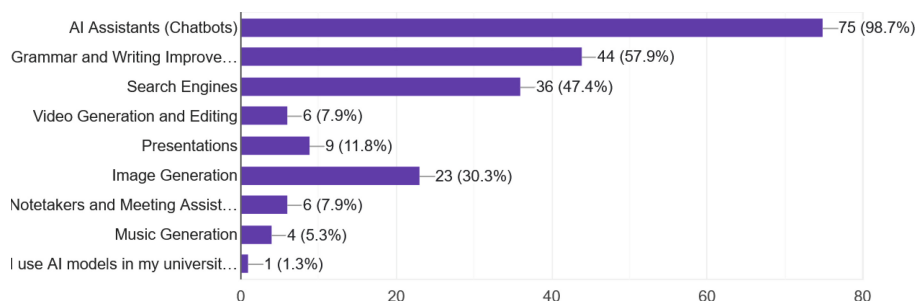


Fig. 1. What types of AI tools do you use?

Students use AI most frequently 2–4 times per week (42.1%), with daily use being the next most common (31.6%) as shown in Fig. 2.

For those students who had used AI Assistants (Chatbots), 96.1% had used ChatGPT, 43.4% Microsoft Copilot, Gemini 34.2%, 25% Claude, 18.3% Deepseek, and others like MathAI, Mistral, Grok in 1–2%. 59 students use AI Assistants because they are free, 35 due to their popularity, and 26 based on recommendations. On the other hand, 28.9% of students would not pay for the usage of AI Assistants (Chatbots), 51.3% would consider it, and almost 20% would agree to pay.

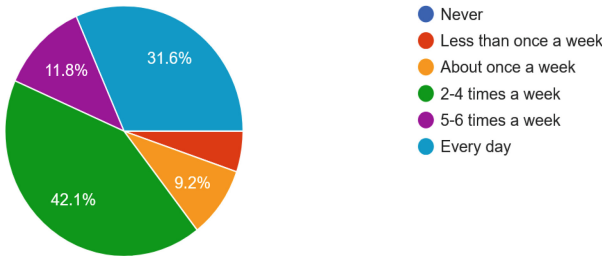


Fig. 2. How frequently do you use AI tools?

Only two students have not used AI tools to support teaching or learning activities, while others use it mostly for solving problems (81.6%), programming (78.9%), teaching (55.3%), writing homework/tasks (39.5%), and writing essays (28.9%).

Open-ended comments relating to how students were motivated to adopt AI in your learning. The following topics were raised: *“It is simpler to use that visiting forums”*, *“It easy to find answers to problems”*, *“Possibility to learn faster, in less annoying way than usually”*, *“It makes everything much faster”*, *“Insufficient information from teachers”*, *“It basically just optimizing my time you spent on learning.”*.

AI tools used in an educational context are: AI Assistants (Chatbots) (97.4%), Grammar and Writing Improvement 52.6%, Search Engines 35.5%, Image Generation 11.8%, Presentations and Notetakers and Meeting Assistants 10.5%, and Video Generation and Editing 3.9%.

46 (60.5%) students consider effectiveness of AI in enhancing student learning outcomes as “Effective” or “Absolutely effective”, while tw students consider it “Ineffective” (see Fig. 3). Further, 59.2% of students think that AI tools help students become more independent learners, whilst 21.1% disagree.

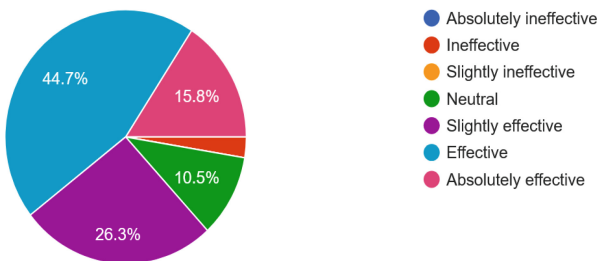


Fig. 3. In your opinion, how effective is AI in enhancing student learning outcomes?

4.3 Satisfaction, Ethics and Confidence

First we asked students if they have used AI to generate complete tasks for homework, and 51.3% of students confirmed that. Next, we asked to range on a 5-point Likert scale with the anchors “Strongly agree” to “Strongly disagree” with the statement “*AI-assisted academic dishonesty is a serious threat to educational integrity*”, and almost 55% agrees with the statement (see Fig. 4).

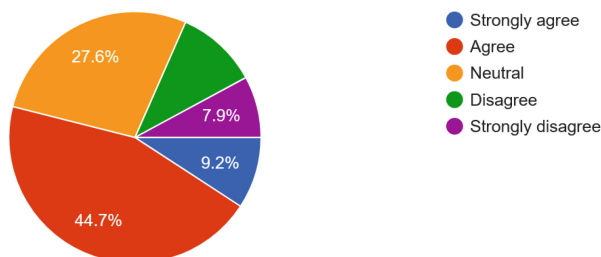


Fig. 4. AI-assisted academic dishonesty is a serious threat to educational integrity.

To 69.7% of students think that students need to be explicitly taught how to use AI ethically.

Students opinion to AI-assisted academic dishonesty and how it should be treated is following:

- It undermines learning and should be strictly penalized (2.6%)
- It’s a growing concern that needs better policies and awareness (21.1%)
- It reflects a shift in how students approach learning and assessment (40.8%)
- It’s not significantly different from other forms of cheating (26.3%)
- It’s inevitable—education systems need to adapt rather than punish (52.6%)
- I don’t see it as a serious issue (21.1%)
- Other (7.8%)

94.7% of students are aware of plagiarism detection software, and 88.2% are familiar with AI detection tools (e.g., GPTZero, Turnitin AI detection). Each term—‘plagiarism detection’ and ‘AI detection’—was unfamiliar to one female student, respectively.

Only 21% of students is aware of guidelines for the use of AI in education at their university, while almost half of students support the implementation formal policies on AI use in schools or universities. 60.5% of students think that AI will replace certain teaching tasks or roles in the future, but 25% disagree.

All students confirmed that they come across to AI answers that are not correct, and 93.4% asked the AI to correct answers that are not correct. To what extent do you trust the outputs generated by AI is shown in Fig. 5. Students trust the outputs generated by AI in the following context:

- I trust AI outputs when verified by other sources (47.4%)

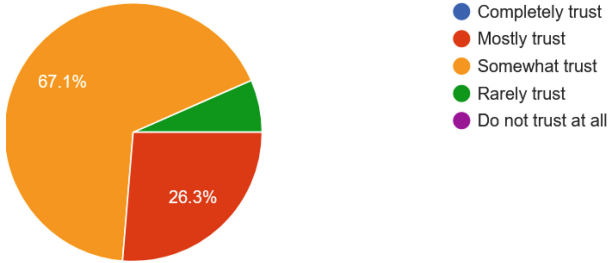


Fig. 5. To what extent do you trust the outputs generated by AI?

- I trust AI for basic tasks but not for critical decisions (59.2%)
- I only trust AI when I understand how it produced the output (68.4%)
- I trust AI depending on the specific tool or platform (9.2%)
- I don’t trust AI-generated outputs without human oversight (11.8%)
- Other (3.9%)

5 Conclusion

In this research, we have identified some of the challenges that secondary and higher education institutions and students may face due to the integration of new technologies into teaching and the pedagogical process. It turns out that the integration of AI, especially large language models (LLMs), represents a new dimension in the teaching of formal methods. In addition to didactic support for practicing concepts, they also allow the generation of structured outputs that can be analyzed, transformed and verified using traditional formal techniques. The research thus confirms the opening of new possibilities for their targeted use, e.g. in analytical and verification tasks.

Collaboration in the development and implementation of AI-based digital solutions helps students, teachers and university management to better understand the challenges and opportunities of digital transformation. This shared experience deepens their commitment to the project goals and supports a culture of continuous improvement within the institutions.

This research thus deepens our understanding of the potential and limitations of large language models (LLMs) and their integration into electronic assessment systems, pedagogical scenarios, and student teaching through GPT-based applications. The results also point to the need for further exploration of the possibilities of using LLMs in analytical and verification tasks, which may open new perspectives in the field of formal methods teaching.

Acknowledgments. This work was supported by the project 030TUKE-4/2023 – “Application of new principles in the education of IT specialists in the field of formal languages and compilers”, granted by the Cultural and Education Grant Agency of the Slovak Ministry of Education; ALFA Labs – trAnsformation pLan For Ai Living Labs, and CZ-1503-05-2425 (Umbrella) – Development of Computational Thinking.

References

1. European education area. <https://education.ec.europa.eu/>. Accessed 8 Feb 2025
2. European framework for digitally competent educational organisations - dig-comporg (2024). https://joint-research-centre.ec.europa.eu/european-framework-digitally-competent-educational-organisations-digcomporg_en. Accessed 8 Feb 2025
3. Selfie - a tool to support learning in the digital age (2024). <https://education.ec.europa.eu/selfie>. Accessed 8 Feb 2025
4. AI in Education (2025). <https://forms.gle/R5UtcyEjE5xTGkmT6>. Accessed 19 July 2008
5. Baidoo-Anu, D., Owusu Ansah, L.: Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *J. AI* **7**(1), 52–62 (2023)
6. Brognoli, P.C., Rodrigues, M.B.: Uso do ChatGPT por estudantes italianos na universidade degli studi di perugia-unipg. *Revista Nova Paideia - Revista Interdisciplinar em Educação e Pesquisa* **7**(1), 332–350 (2025). <https://doi.org/10.36732/riep.v7i1.421>, <https://ojs.novapaideia.org/index.php/RIEP/article/view/421>
7. Carolus, A., Koch, M.J., Straka, S., Latoschik, M.E., Wienrich, C.: Mails - meta AI literacy scale: development and testing of an AI literacy questionnaire based on well-founded competency models and psychological changeand meta-competencies. *Comput. Hum. Behav. Artif. Hum.* **1**(2), 100014 (2023). <https://doi.org/10.1016/j.chbah.2023.100014>, <https://www.sciencedirect.com/science/article/pii/S2949882123000142>
8. Chen, L., Chen, P., Lin, Z.: Artificial intelligence in education: a review. *IEEE Access* **8**, 75264–75278 (2020). <https://doi.org/10.1109/ACCESS.2020.2988510>
9. Çildir, M.: Collaborating with ChatGPT: an exploratory study of German language learning. *Alman Dili ve Kültürü Araştırmaları Dergisi* **5**(2), 63–76 (2023)
10. Crawford, J., Cowling, M., Allen, K.A.: Leadership is needed for ethical ChatGPT: character, assessment, and learning using artificial intelligence (AI). *J. Univ. Teach. Learn. Pract.* **20**(3) (2023)
11. Fauzi, F., Tuhuteru, L., Sampe, F., Ausat, A.M.A., Hatta, H.R.: Analysing the role of ChatGPT in improving student productivity in higher education. *joe* **5**(4), 14886–14891 (2023)
12. Guo, S., Shi, L., Zhai, X.: Validating an instrument for teachers' acceptance of artificial intelligence in education. [arXiv:2406.10506](https://arxiv.org/abs/2406.10506) (2024)
13. Ilić, J., Ivanović, M., Klačnja-Milićević, A.: The impact of ChatGPT on student learning experience in higher stem education: a systematic literature review. In: 2024 21st International Conference on Information Technology Based Higher Education and Training (ITHET), pp. 1–9 (2024). <https://doi.org/10.1109/ITHET61869.2024.10837649>
14. Ivanović, M., Klačnja-Milićević, A., Paprzycki, M., Ganzha, M., Bădică, C., Bădică, A., Jain, L.C.: Current Trends in AI-Based Educational Processes—An Overview, pp. 1–15. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-04662-9_1
15. Lamas, P., Arnab, S.: Power to the teachers: an exploratory review on artificial intelligence in education. *Information* **13**(1) (2022). <https://doi.org/10.3390/info13010014>
16. Mallik, S., Gangopadhyay, A.: Proactive and reactive engagement of artificial intelligence methods for education: a review. *Front. Artif. Intell.* **6** (2023). <https://doi.org/10.3389/frai.2023.1151391>

17. Mike, N., et al.: The impact of ChatGPT on student performance in higher education. In: *Proceedings of Ongoing Research, Practitioners, Posters, Workshops, and Projects of the International Conference EGOV-CeDEM-ePart 2024*, vol. 3737. Ghent/Leuven, Belgium (2024)
18. Ng, D.T.K., Leung, J.K.L., Chu, S.K.W., Qiao, M.S.: Conceptualizing AI literacy: an exploratory review. *Comput. Educ. Artif. Intell.* **2**, 100041 (2021). <https://doi.org/10.1016/j.caeai.2021.100041>, <https://www.sciencedirect.com/science/article/pii/S2666920X21000357>
19. Nieto-Jalil, J.M., Isrrael Tec Chim, A., Seuret-Jimenez, D., Martínez Huerta, J.M.: The impact of GPT, experiential learning, and reinforcement methods on complex problem solving. In: *2025 IEEE Engineering Education World Conference (EDU-NINE)*, pp. 1–6 (2025). <https://doi.org/10.1109/EDUNINE62377.2025.10981377>
20. Owoc, M.L., Sawicka, A., Weichbroth, P.: Artificial intelligence technologies in education: benefits, challenges and strategies of implementation. In: Owoc, M.L., Pondel, M. (eds.) *Artificial Intelligence for Knowledge Management*, pp. 37–58. Springer International Publishing, Cham (2021)
21. Padilla, J.R.C., Montefalcon, M.D.L., Hernandez, A.A.: Language AI in programming: a case study of ChatGPT in higher education using natural language processing. In: *2023 IEEE 11th Conference on Systems, Process & Control (ICSPC)*, pp. 276–281 (2023). <https://doi.org/10.1109/ICSPC59664.2023.10420194>
22. Roca, M.D.L., Chan, M.M., Garcia-Cabot, A., Garcia-Lopez, E., Amado-Salvatierra, H.: The impact of a chatbot working as an assistant in a course for supporting student learning and engagement. *Comput. Appl.n Eng. Educ.* **7**(32), e22750 (2024). <https://doi.org/10.1002/cae.22750>, <https://ojs.novapaideia.org/index.php/RIEP/article/view/421>
23. Silva, C.A.G.D., Ramos, F.N., de Moraes, R.V., Santos, E.L.D.: ChatGPT: challenges and benefits in software programming for higher education. *Sustainability* **16**(3), 1245 (2024)
24. Taherdoost, H.: Data collection methods and tools for research; a step-by-step guide to choose data collection technique for academic and business research projects. *Int. J. Acad. Res. Manag. (IJARM)* **10**(1), 10–38 (2021). <https://hal.science/hal-03741847>
25. Thorp, H.H.: ChatGPT is fun, but not an author. *Science* **379**(6630), 313–313 (2023). <https://doi.org/10.1126/science.adg7879>
26. Wu, Y.: Integrating generative AI in education: how ChatGPT brings challenges for future learning and teaching. *J. Adv. Res. Educ.* **2**(4), 6–10 (2023). <https://doi.org/10.56397/JARE.2023.07.02>
27. Zheng, Y.: ChatGPT for teaching and learning: an experience from data science education. In: *SIGITE 2023, Proceedings of the 24th Annual Conference on Information Technology Education*, pp. 66–72. Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3585059.3611431>



Impact of Gender Bias in the Output of AI Language Models on Heavy Users

Sarah Hosell^(✉)

Institute of Computer Science, Hochschule RuhrWest, University of
Applies Science, Mulheim, Germany
sarah.hosell@hs-ruhrwest.de

Abstract. Research highlights how AI-generated narratives often reflect gender stereotypes, associating feminine characters with family and emotions, and masculine characters with politics and war. The key Goals are to investigate how AI models' gender bias influences perceptions and decisions of frequent users and also to explore ways to adapt findings from prior studies to improve AI interactions and reduce the impact of bias in outputs. Students input five prompts into the AI, with outputs varying between generic masculine (experimental group) and gender-neutral/feminine-masculine forms (control group). They respond to questions about their desired childhood career (categorized by gendered or neutral terms) and rate the difficulty of gendered and neutral professions on a scale of 1–5. Responses are evaluated by gender and group (experimental vs. control). Cluster analysis via SPSS will create user persona (personalized fictional character).

The experiment has received approval from the ethics committee, because participants remain unaware of the study's purpose to avoid biased results.

Keywords: Gender Bias · Output of AI Language · Relationship between Gender Bias and Job Assessment · Relationship between gender bias and difficulty assessment of a job

1 Introduction

In stories generated by AI language models, masculine and feminine characters are placed in different topics: Feminine characters are more commonly associated with discussions about family, emotions, and body parts, whereas masculine characters are more frequently linked to topics such as politics, war, sports, and crime [1]. Generated narratives exhibit multiple gender stereotypes, which can surface even when the prompts lack explicit gender indicators or stereotype-related themes [1].

And Science promotes: A bibliometric study analyzed published articles on gender bias, revealing that they receive less funding and are published in journals with lower Impact Factors compared to articles on similar forms of social discrimination [2].

At the same time, it is clear that what is implied does not necessarily mean it has been thoroughly considered. Two classroom experiments with 591 primary school students from two different linguistic backgrounds (Dutch or German) showed that the

presentation of occupational titles in pair forms (e.g., Ingenieurinnen und Ingenieure, female and male engineers), rather than in generic masculine forms (Ingenieure, plural for engineers), boosted children's self-efficacy with regard to traditionally male occupations, with the effect fully being mediated by perceptions that the jobs are not as difficult as gender stereotypes suggest [3]. This means that what is meant is not what is thought of.

Various strategies have been proposed to address these sources of bias, such as dataset augmentation, bias-aware algorithms, and incorporating user feedback mechanisms [4].

1.1 Terminological Foundations

In this subchapter, key terms of the research are defined. Relevant literature is reviewed to provide definitions, which are then discussed, and the meanings most relevant to this study are extracted.

Gender Bias. In 1955, the concept of gender role or gender identity was introduced to discuss the discrepancy between expected and actual behavior [5] of such people. In this context, the anglicism gender is usually translated as “social gender” in the German-speaking world and is used primarily for analytical categorization. In this research work the concept gender is also associated as social status, things that a person says or does to identify themselves as someone who has the status of a man or boy, a woman or girl. Bias is understood here as systematic distortion.

In scientific research, gender bias refers to a distortion of reality through formulations, mental assumptions or statistical errors that leads to false representations of the actual gender-specific relationships [6]. Gender bias thus refers to systematic distortion effects that arise from gender-related stereotyping and prejudice in the text-based outputs of LLMs. Gender bias in text editions usually consists of overgeneralization. This means that the experiences or perspectives of only one gender are adopted and applied to both genders.

Output of AI Language Models. AI language model output refers to texts, responses, or generated content produced by artificial intelligence systems that process and generate human-like language. In this research work, only editions in text form are taken into account.

Heavy Users. Heavy Users are individuals who engage with a product, service, or technology at a significantly higher frequency or intensity than the average user [7]. In the context of digital platforms and AI technologies, heavy users are those who rely on these tools extensively in their daily activities, whether for work, education, or personal use [8]. In the context of using AI language models in this research work, this means: use more than several times a week and involvement in activities such as research, brainstorming, improving texts, translating texts, and analyzing data.

1.2 Usage of AI Systems

The use of AI in the general population is demonstrated through various studies: 35.5% use ChatGPT at least several times a week. In a professional context, ChatGPT is most frequently used for research, brainstorming, data analysis and content summarization [9]. The 25- to 35-year-old people [9] represents the largest user group.

Around 47% [10] of respondents currently use ChatGPT to write texts or for creative writing. 43% [10] use ChatGPT as a search engine to obtain information.

The group of pupils (84%) and students (75%) uses ChatGPT the most [11]. Young people from Germany say they are most regularly in contact with artificial intelligence in the form of voice assistants [12]. Young people least often report using artificial intelligence [12], for example, robotics. Around 80% [12] said they had never used an AI of this kind in everyday life.

Private use is dominated by ChatGPT at 52% [13]. Another 37% [13] use smart speakers, such as Alexa or Google Assistant, in everyday life. Older people (65 +) are more skeptical about AI, while younger people (<35) are more likely to experiment with LLMs like ChatGPT and actively use them for education and work [14]. Differences exist not only in age but also in gender with regard to the use of LLMs.

Otis et. al [15] found out, that gender gap in generative AI usage holds across nearly all regions, sectors, and occupations. It is documented that this gap persists even when equal access to this new technology is provided. In relative to the share of men who use generative AI, the share of women is often about 10% to 40% smaller. For example women are much less likely to use generative AI than men in occupations ranging from postdocs across the world (21%) to business owners in the US, Australia, the UK, and Canada (11%) to college students in the US and Sweden (25% and 31%). An exception is a Boston Consulting Group (BCG) study showing that women are 3% [16] more likely to use generative AI than men in a convenience sample of 6,558 tech employees. This positive difference is mainly due to women also in leading technical positions in the population surveyed.

This suggests that going beyond users' socio-demographic data could provide deeper insights. More relevant factors include usage motives and the social class of high-frequency users.

Kacperski et. al [17] postulate, that lower age and higher education and also great affinity for technology correlate with the use of ChatGPT. Likewise, the use of various social media is positively correlated with ChatGPT activity. In terms of political variables, ChatGPT activity was associated with political knowledge, political self-efficacy, and online political action. However, online political debating and political self-efficacy showed a negative association.

A significant share of those who primarily use LLMs for school and work are highly educated individuals. This suggests that a certain level of cognitive reflection and intellectual oversight of AI output is essential. At the same time, it shows the danger of AI-generated expenditures if they are adopted without reflection or if it is not even recognized that they were created by an AI.

2 Methodology

2.1 Research Question

This Gender Bias in the Output of AI Language models is still there. So the research questions are:

- What impact does gender bias in the output of AI speech models have on high-frequently users or who use these systems frequently?
- Can the results of Vervecken and Hannover be easily transferred to the everyday use of AI language models?

The special socio-demographic and psycho-graphic characteristics respectively, motivation to use and behavioral characteristics of these heavy users must be taken into account.

2.2 Qualitative Analysis by Creating User Persona

Based on the findings from Sect. 1.2, fictitious personas were developed that represent “typical” users. A persona also user persona [18] in user-centered design and marketing is a personalized fictional character created to represent a potential end user. The user persona were initially developed and subsequently reviewed and refined through qualitative interviews. The final design of the persona is as follows. Tim, Anna, Charlotte and Oliver do not exist, so anonymization is not necessary. Gender parity was explicitly insisted upon here, even though the typical user is male. However, based on the BCG study, a priority can be justified (Table 1).

2.3 Research Approach

To examine the impact of gender bias in the output of AI speech models on high-frequency users, an experiment involving an experimental group and a control group will be conducted.

The experiment is designed as follows:

Students aged 18 to 30 will be invited to participate in a laboratory experiment. Computer science students were selected here because this selection corresponds to the frequent users described in Sect. 1.2: mostly male, well educated, tech-savvy. These students are asked about their character traits and assigned to the corresponding personas. These participants will be required to input five predefined prompts into the AI system and read the generated responses. The experimental group receives inputs and outputs in the generic masculine. The control group receives inputs and outputs in neutral or in the feminine and masculine form.

At the same time, the output from ChatGPT is captured and the reading time is stopped. Following this, they will answer a few of open-ended questions, with and without no predefined response options provided.

First, the students are asked about their desired career as a child. It is recorded whether they indicated a male, female or a neutral form of the profession.

Table 1. Table fictional user persona

Fictional persona	Socio-demographic	psychographic characteristics	motivation to use	behavioral characteristics
Person Tim	year of birth 1998 university degree in business administration high affinity for technology Profession: Consultant	uses ChatGPT to simplify tasks and solve them more quick Strives for quick answers	Time saving: Efficient support with tasks or projects most common use: email answers written by ChatGPT everyday use, often, quick	makes targeted and precise inquiries time is money Use AI language models as a personal assistant
Person Anna	year of birth 2003 biology student affinity for technology	uses ChatGPT for inspiration and creative projects Thirst for knowledge	uses ChatGPT as a partner for brainstorming or reflection weekly use, not often, but longer	Experimenting: Fascination with the tech-nology and the possibilities offered by AI Access to knowledge: answers to questions or complex topics in simple language
Person Charlotte	year of birth 1989 Doctoral university degree in International Management high affinity for technology Profession: Analytics department at Google Germany	uses AI language models to prepare analyses, summaries and protocols runs first analysis via Ki, but checks everything again herself	most common use: Support in evaluating large amounts of data and identifying trends everyday use	at home worldwide, Openmind Use AI language models as a tool, but never without control
Person Oliver	year of birth 2004 student of business informatics Very high affinity for technology	uses AI language models (mostly ChatGPT) to generate texts for seminar papers, for example uses AI language models (DeepL Write) to improve self-written texts use Adobe AI images generator (Adobe firefly)	save time writing text is very time-consuming, he prefers to let the AI do it have free images created to use in websites created websites as a part-time job	takes a lot of time to go out with friends want to get through your studies quickly and comfortably, He thinks it is so important to be able to formulate texts very well

Secondly, the field of work of a typically female and a typically male and also typically neutral profession is read out in neutral language. The students from both groups are asked to rate the level of difficulty of these professions on a scale. They are also asked whether they feel confident in pursuing these careers. This self-esteem is a relatively stable personal characteristic.

The SEKJ Self-Esteem Inventory is used to measure this confidence in pursuing their chosen career—presented here in this test setting in a male or neutral form. The SEKJ is divided into self-esteem level (SWH), self-esteem stability (SWS), and self-esteem contingency (SWK). Each block contains 10 or more items with a 5-point response

format (1 = not at all true, 5 = exactly true) that capture the respondents' statements [19].

- (SWH) **Self-esteem level:** Extent/level of a person's global self-esteem
- (SWS) **Self-esteem stability:** The totality of a person's perceived fluctuation and fragility of their own worth (opposite: stable, robust self-esteem)
- (SWK) **Self-esteem contingency:** The dependence of self-esteem on achieving certain self- or externally set standards – here: in the area of academic competence and performance (opposite: independent, “unconditional” self-esteem) [19]

The SEKJ is primarily needed here to compare a possible bias in self-confidence regarding a profession with one's own self-esteem. This way, a bias rooted in personality based on self-esteem can be prevented.

Due to the length of the self-esteem inventory, appropriate items are selected from each block for this research.

2.4 Summary of the Relationships and Derivation of the Hypotheses

Derivation of the thesis and hypothesis:

H_1 The younger the users, the more they are affected by the effects of gender bias in AI outputs.

H_2 The higher the level of education, the higher the ability to compensate for gender bias.

H_3 The higher social media usage, the more the impact of gender bias in AI output

H_4 Heavy users are more affected by the effects of gender bias in AI outputs.

H_5 Male users are more affected by the effects of gender bias in AI outputs.

H_6 Heavy female users in tech jobs can better reflect the effects of gender bias in AI outputs.

H_7 AI as search engine-users are more affected by the effects of gender bias in AI outputs.

2.5 Derivation of the Questionnaire and Experiment

In the subsequent phase, both the questionnaire and the experimental design were developed from the framework. A pretest involving 3 participants only for testing the experiment was then conducted to evaluate the clarity and feasibility of the materials and procedures. In the main experiment, the adjusted population comprised 31 subjects. Based on insights gained from the pretest, several adjustments were made to the experimental setup:

- Since the group of surveyed students was relatively homogeneous—particularly in terms of their current highest educational qualification—we additionally collected information on the highest educational qualification of their parents.

- Most students indicated that they used ChatGPT on a daily basis. Therefore, if a participant selected “daily” for frequency of use, a follow-up question was added to assess the specific number of daily interactions.
- The ChatGPT experiment revealed that AI-generated responses were inconsistently gendered—some outputs displayed noticeable gender bias, while others did not. As a result, it became necessary to pre-generate all AI responses in a controlled lab environment prior to the field experiment. This approach ensures consistency and allows for the creation of both experimental and control groups.

In the following, the experiment is set up to record the effect, impacts of gender bias as well as the possibility of reflecting on and compensating for gender bias. The experiment is carried out in three steps:

- first, the time required by users to read texts with and without gender formulations is measured during the experiment.
- second, the effects and impacts of Gender Bias are investigated and
- third, the possibility of reflecting and compensating gender bias in AI-outputs is being researched.

Inspired by the experiment by Vervecken and Hannover, “Effects of gender-fair job descriptions on children’s perceptions of job status”, participants in the pretest were asked about their childhood dream job. Their response was used to generate a descriptive text about the profession using ChatGPT. This description was then read aloud to the participants. They were subsequently asked whether they could imagine themselves working in this profession today. The presence of gender bias in this live test with ChatGPT proved to be unpredictable. Moreover, gender bias was often quite obvious for instance, through the exclusive use of either the male or female form. In many cases, the participants recognized this immediately.

As a result, the experiment was modified in two steps: Participants were still asked about their childhood dream job, and a live AI-generated image of the profession was created. They were then asked whether the image matched their typical idea of that profession. Secondly, prior to conducting the experiment, a fictional product a fitness watch was created using AI, along with a product description. After generating several versions, two descriptions were selected as Version A and Version B, respectively, for use in the study.

All participants were interviewed personally, and the sample was selected randomly. The final adjusted sample comprised 31 participants (Figs. 1 and 2).

PULSE Motion

Your companion for everyday balance

Whether during yoga, daily routines, or between lectures and relaxation – the PULSE Calm supports users who care about their well-being.

With features like sleep and stress analysis, cycle tracking, gentle activity reminders, and breathing exercises, it accompanies athletes and wearers throughout the day. Its elegant, lightweight design fits comfortably on the wrist – available in various colors.

The intuitive operation and clear display make it ideal for anyone who wants to use smart technology without technical clutter.

PULSE Calm – for users who want to combine health and style.



Fig. 1. Experiment version in the generic masculine

PULSE Motion

Engineered for ambitious athletes

The PULSE Motion is the perfect training companion for athletes who want to keep a close eye on their performance. With features like GPS tracking, VO₂max measurement, interval timer, and cadence analysis, this watch offers everything an athlete needs for effective training.

Its scratch-resistant casing and 12-day battery life make it especially durable – whether during strength training, running, or mountain biking.

A structured training mode, automatic recovery analysis, and personalized goal setting help wearers stay on track and push toward their fitness goals with focus and consistency.

PULSE Motion – for athletes who aim higher than average.



Fig. 2. Experiment version in feminine form

Participants were randomly assigned either Version A or Version B of the text. Subsequently, they were asked four questions to evaluate the text: to assess...

- impact, “Do you find the text appropriate?”;
- effects, “Do you feel addressed by the text?”;
- compensate, “Did anything in the text particularly catch your attention?”; and
- reflect, “How should the text be modified to better suit your preferences?”

The responses yielded both quantitative and qualitative data. The face-to-face setting also allowed for follow-up questions and in-depth discussion.

3 Results

The following chapter conducts a quantitative analysis and evaluation of the hypotheses, as well as a qualitative analysis using Mayring's content analysis.

3.1 Quantitative Analysis and Evaluation of the Hypotheses

Firstly, it was found that texts using both feminine and masculine forms disrupt reading flow. Texts written in the generic masculine are the most fluent, the test persons needed an average of 11.94 s. Those in neutral form fall somewhere in between and need on average 20.44 s.

The classification into personas was only partially successful. The usage patterns of AI in everyday school and work life have progressed significantly, so the motivation and behavioral characteristics from the personas applied to almost all test subjects.

When asked about their childhood career aspirations, the vast majority of respondents regardless of gender used the masculine form of the profession.

Evaluation of the hypotheses:

H_1 *The younger the users, the more they are affected by the effects of gender bias in AI outputs.*

Hypothesis 1 must be falsified based on the available data, Pearson Chi-Square with 16.439 and Asymptotic significance 2-sided 0.037. Since the age of the respondents was rather homogeneous, a representative survey based on age should be conducted.

H_2 *The higher the level of education, the higher the ability to compensate for gender bias.*

Hypothesis 2 must be falsified based on the available data, Pearson Chi-Square with 2.399 and Asymptotic significance 2-sided 0.663. However, it is striking that the cases in which the ability to compensate for gender bias was discovered are the same cases whose parents had the highest academic degree.

H_3 *The higher social media usage, the more the impact of gender bias in AI output.*

Hypothesis 3 can be verified at a low significance level, Pearson Chi-Square with 24.898 and Asymptotic significance 2-sided 0.002. The cross tabs output shows clear a correlation. Social media consumption correlates with the effect of gender bias.

H_4 *Heavy users are more affected by the effects of gender bias in AI outputs.*

In hypotheses 4, the increased use of ChatGPT also shows an increased susceptibility of users to the effects of gender bias. The ANOVA shows a weak connection in regression with T at 0.727, sig 0.473.

H_5 *Male users are more affected by the effects of gender bias in AI outputs.*

The results of the cross-tabulation for H5 show that an equal number of male and female participants did not notice the gender bias effect. Therefore, the hypothesis is rejected. This corresponds with the low level of statistical significance at 0.118 and Pearson Chi-Square 4.277. However, it is noteworthy that twice as many women as men

perceived the gender bias at least indirectly, in the form of a sense that the text felt inappropriate.

H_6 *This hypothesis could not be evaluated due to a lack of women in tech jobs.*

H_7 *AI as search engine-users are more affected by the effects of gender bias in AI outputs.*

Hypothesis 7 is verified due to the correlation between the number of subjects who use AI as a search engine and at the same time are more affected by the effect of gender bias. The low significance level at 0.227 is due to the small number of cases.

3.2 Qualitative Analysis Using Content Analysis According to Mayring

The evaluation of the qualitative data collected during the discussion and questions in the experiment is carried out using qualitative content analysis according to Mayring.

The qualitative data collected during the discussions and follow-up questions were analyzed using qualitative content analysis according to Mayring. The participants’ responses to the questions related to “effect,” “compensate,” and “reflect” in the experiment were transcribed and systematically analyzed. It is important to note that no quantification of responses was conducted in this process.

As part of the experiment and the question “Do you feel addressed?” (effects), the following response categories were extracted.

for the <i>masculine</i> form:	for the <i>feminine</i> form:
· Appropriate	· Feeling addressed
· Advertising language	· Soft tone
· Data/facts	· Too general
· Not engaging	

In response to the question “Did you notice anything in the text?” (compensate), the following categories were identified.

for the <i>masculine</i> form:	for the <i>feminine</i> form
· Insufficient	· Gender-related comment
· Needs more information	· Impersonal
· Nothing noticed	· Needs more technical information

Regarding the question “How should the text be written to better suit you?” (reflect), the following categories were identified for both masculine and feminine forms:

- Text feels inappropriate
- Nothing noticed

Particularly the very brief and limited responses to the “compensate” and “reflect” questions suggest that participants were often able to express a general sense that a text felt inappropriate or not appealing, but were not able to clearly verbalize what caused this impression. This highlights the importance of using gender-appropriate and inclusive language to ensure that readers feel explicitly addressed. The key takeaway here is: being included implicitly is not the same as being explicitly considered.

4 Conclusions and Implications

The implications are discussed based on the research questions.

What impact does gender bias in the output of AI speech models have on high-frequency users or who use these systems frequently?

One effect is a disruption of the reading flow when using gender-neutral text forms and especially when using feminine and masculine text forms at the same time.

The findings indicate that high-frequency users are less likely to detect potential gender bias in the textual output of AI language models. This effect is particularly pronounced among individuals with lower educational attainment, who often lack a sensitivity to linguistic manifestations of gender bias. In this context, increased levels of education appear to foster a heightened awareness—supporting the notion that educational exposure plays a key role in bias recognition and critical language competence.

Simultaneously, extensive social media use correlates with a reduced likelihood of recognizing gender-biased language, suggesting that habitual engagement with algorithmically curated content may desensitize users to subtle forms of bias.

Notably, female participants were more likely to report a sense of discomfort or inappropriateness in response to gender-biased language. This observation may reflect a decreased sense of identification or inclusivity when gendered language fails to align with inclusive linguistic norms.

The impact of gender bias was especially evident among high-frequency users who rely on ChatGPT as a search engine or everyday assistant. This highlights the dual-edged potential of AI language models: while they offer considerable utility and accessibility, their uncritical integration into daily life also introduces risks of reinforcing existing biases.

Strikingly, only two participants were able to explicitly identify, articulate, and compensate for the presence of gender bias in the texts. Several others expressed a vague sense of linguistic inappropriateness, yet lacked the conceptual or linguistic tools to formulate a response. These findings suggest that effective compensation for bias requires not only awareness but also the ability to linguistically frame the issue—a competence likely tied to education and critical language literacy.

Can the results of Vervecken and Hannover be easily transferred to the everyday use of AI language models?

The results of Vervecken and Hannover also prove to be reliable for adult frequent users in this experiment. When the text describing typically male professions was presented in the masculine form, the participants responded with a higher rating of the difficulty level. In contrast to the preschool children, the female participants rated the difficulty level even higher.

The practical implications are: If self-assessment in a specific group of people can be improved by using both forms, male and female, and if the increase in self-confidence is accompanied by an increase in attempts, the general introduction of both forms is recommended.

This clearly shows that the use of the feminine and masculine forms, or a suitable form, is particularly important in giving the reader the feeling of being included. Here, the following applies: “I meant” does not mean “I thought”.

Acknowledgments. Thank you to the ethics committee and to all participants to my assistant. The experiment will be evaluated by the ethics committee of the Ruhr West University of Applied Sciences before it is carried out, as the subjects cannot be made aware of the purpose of the experiment before it is carried out. This would falsify the results.

Disclosure of Interests. The author has no conflicts of interest to disclose that are relevant to the content of this article.

References

1. Lucy, L., Bamman, D.: Gender and Representation Bias in GPT-3 Generated Stories. In: Proceedings of the Third Workshop on Narrative Understanding, pp 48–55. Association for Computational Linguistics, Virtual (2021)
2. Cislak, A., Formanowicz, M., Saguy, T.: Bias against research on gender bias. *Scientometrics* **115**, 189–200 (2018). <https://doi.org/10.1007/s11192-018-2667-0>
3. Vervecken, D., Hannover, B.: Yes I can! Effects of gender fair job descriptions on children’s perceptions of job status, job difficulty, and vocational self-efficacy. *Social Psychology* **46**(2), 76–92 (2015). <https://doi.org/10.1027/1864-9335/a000229>
4. Ferrara, E.: Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci.* **6**(1), 3 (2024). <https://doi.org/10.3390/sci6010003>
5. Money, J., Hampson, J.G., Hampson, J.: An Examination of Some Basic Sexual Concepts: The Evidence of Human Hermaphroditism. In: Bulletin of the Johns Hopkins Hospital. Band 97, Nr. 4, 1. Oktober 1955, S. 301–319 (englisch; PMID 13260820) (1955)
6. Eichler, M., Gustafson, D. L., Pompetzki, M.: Moving toward equality: Ottawa: Health Canada. In: Gustafson, D. L. (ed.) *Toward Inclusionary Practices in the Education of Nurses: A Critique of Transcultural Nursing Theory*. The Alberta Journal of Educational Research **45**(4), 468–470 (1999)
7. Twedt, D.W.: How Important to Marketing Strategy is the “Heavy User”? *Journal of Marketing* **28**(1), 71–73 (1964). <https://doi.org/10.1177/002224296402800116> (Original work published 1964)
8. Emmanouilides, C., Hammond, K.: Internet usage: Predictors of Active Users and Frequency of Use. *J. Interact. Mark.* **14**(2), 17–32 (2000). [https://doi.org/10.1002/\(SICI\)1520-6653\(200014\)14:2%3c17::AID-DIR2%3e3.0.CO;2-E](https://doi.org/10.1002/(SICI)1520-6653(200014)14:2%3c17::AID-DIR2%3e3.0.CO;2-E)
9. Gedankenfabrik: Wofür wird ChatGPT in Ihrem Unternehmen genutzt?. Statista. Statista GmbH (2023). Last access: 13. Januar 2025. <https://de-statista-com.hs-ruhrwest.idm.oclc.org/statistik/daten/studie/1401309/umfrage/chatgpt-nutzung-in-unternehmen/>
10. Initiative D21. Gründe für die Nutzung von ChatGPT in Deutschland im Jahr 2023. Statista. Statista GmbH (2024). Last access: 13. Januar 2025. <https://de-statista-com.hs-ruhrwest.idm.oclc.org/statistik/daten/studie/1465935/umfrage/gruende-zur-nutzung-von-chatgpt-in-deutschland/>

11. Evergreen Media: Haben Sie ChatGPT bereits verwendet? Statista GmbH (2024). Last access: 13. Januar 2025. <https://de-statista-com.hs-ruhrwest.idm.oclc.org/statistik/daten/studie/1464401/umfrage/umfrage-zur-nutzung-von-chatgpt-nach-berufsgruppen/>
12. BLM: Häufigkeit der Nutzung verschiedener KI-Anwendungen von Jugendlichen in Deutschland im Jahr 2023. Statista. Statista GmbH (2024). Last access: 13. Januar 2025. <https://de-statista-com.hs-ruhrwest.idm.oclc.org/statistik/daten/studie/1490561/umfrage/haeufigkeit-nutzung-verschiedener-ki-anwendungen-bei-jugendlichen/>
13. VIER: Welche KI-basierte Lösungen nutzen Sie?. In Statista (2024). Last access am 13. Januar 2025. <https://de-statista-com.hs-ruhrwest.idm.oclc.org/statistik/daten/studie/1535310/umfrage/private-nutzung-von-ki/>
14. TÜV Verband, TÜV ChatGPT Studie 2024 (2024). Last access am 27. März 2025. <https://www.tuev-verband.de/pressemitteilungen/zwei-jahre-chatgpt>
15. Otis, N.G., Delecourt, S., Cranney, K., Koning, R.: Global Evidence on Gender Gaps and Generative AI. Harvard Business School (2024)
16. Barisano, M., Neveen, A., Dahik, A., et al.: Women Leaders in Tech Are Paving the Way in GenAI (2024). Last access: 27. March 2025. <https://www.bcg.com/publications/2024/women-leaders-in-tech-are-paving-the-way-in-genai>
17. Kacperski, C., Bonnay, D., Kulshrestha, J.: Characteristics of ChatGPT users from Germany: implications for the digital divide from web tracking data (2023). <https://doi.org/10.48550/arXiv.2309.02142>
18. Lidwell, W., Holden, K., Butler, J.: Universal Principles of Design. Rockport Publishers, p. 182 (2010). ISBN 978-1-61058-065-6
19. Schöne, C., Stiensmeier-Pelster, J.: SEKJ: Selbstwertinventar für Kinder und Jugendliche. Hogrefe Verlag, Göttingen (2016)

Education: Learning and Literacy in the AI Era



Students' Perception of the Integration of GenAI in Academic Paper Assignment Preparation

Katerina Zdravkova^(✉)  and Bojan Ilijoski 

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje,
Skopje, North Macedonia

{katerina.zdravkova,bojan.ilijoski}@finki.ukim.mk

Abstract. The rapid rise of generative artificial intelligence (GenAI), particularly large language models (LLMs), has transformed education, prompting educators to explore ways of integrating these tools into academic workflows. While existing literature discusses their potential and challenges, little empirical research has examined students' experiences with GenAI integration in university settings. This study presents a case from a Computer Ethics course at the Faculty of Computer Science and Engineering, where students completed their assignment implementing three writing types: fully student-written, fully AI-generated, and hybrid. A post-course, self-administered survey was conducted with 43 students, comprising 25 questions across six thematic clusters including: productivity, GenAI usage, assignment process, and ethical concerns. They combined both quantitative and qualitative components. Findings indicate that 62.79% of students were satisfied with the structured GenAI integration, citing increased efficiency and support for learning. Most used LLMs for initial idea generation and content refinement while verifying outputs with reliable sources. The analysis of open responses revealed that 41.67% had a positive experience, while concerns centered on unclear instructions, hallucinated content, and fears of plagiarism. A significant majority (72.09%) recognized ethical risks related to over-reliance on GenAI. The study highlights the need for clear instructional support, ethical frameworks, and flexible, student-centered implementation strategies. It offers practical insights for integrating GenAI in higher education and lays groundwork for future research on long-term impacts and disciplinary differences.

Keywords: Generative AI · AI ethics · higher education

1 Introduction

Following the emergence of ChatGPT and its remarkable success, generative artificial intelligence (GenAI) has rapidly become an indispensable and trusted collaborator across a wide range of disciplines, including education, science, technology, healthcare, finance, the arts, engineering, law, and social sciences [1]. From revolutionizing research and data analysis to enhancing creative expression and streamlining decision-making processes, GenAI is transforming industries and reshaping the way of problem-solving and innovation on a global scale [2].

Large language models (LLMs) have significantly reshaped the educational process by transforming content creation, personalized learning, and student engagement [3]. An extensive scoping literature review [4] revealed that teachers enthusiastically started incorporating GenAI tools in the academic setting by providing teaching, research and writing assistance and enabling personalized learning and education. Many students believe that GenAI encouraged their initial directions and inspiration for research, as well as a focused presentation of key elements of the work [5].

Apart from the exciting opportunities, all previously cited papers [1–5] stress that the integration of GenAI brings with it many ethical challenges that must be carefully considered. They embrace issues of fairness and bias to concerns about privacy, academic integrity, the potential impact on creativity and critical thinking skills, and even the spread of fake news due to inaccurate or misleading GenAI responses, often referred to as “hallucinations” [6]. Another problem is the apparent lack of a regulatory framework for GenAI, leaving educational institutions to define their own rules for integrating GenAI into schooling, teaching, and learning [7, 8].

Beyond these concerns, there is a growing worry about the transparency and explainability of GenAI systems, making it difficult for users to understand how decisions or outputs are generated [9]. This opacity can erode trust and accountability, particularly when GenAI is used to assess student work or support decision-making processes [10]. Additionally, the large-scale data requirements of GenAI models raise questions about data ownership and consent, especially when personal or student-generated content is used for training purposes [11].

The digital divide also presents an ethical dilemma: while some learners benefit from cutting-edge AI tools, others without access to necessary infrastructure or training may fall further behind, exacerbating existing educational inequalities [12]. Finally, the potential for GenAI to reinforce stereotypes or perpetuate harmful cultural assumptions highlights the need for inclusive and context-sensitive design practices [13].

While significant attention has been given to the ethical and legal challenges surrounding the use of GenAI, relatively little research has focused on its practical integration into academic environments, even though many universities actively encourage its adoption [14]. Two studies: a joint experiment involving the use of large language models (LLMs) for academic writing and computer programming [15], and an initiative to integrate chatbots into classroom settings [16] reveal that, despite frequent personal use of LLMs, most students did not rely heavily on these tools when preparing their assignments [15, 16].

The paper proceeds by outlining the motivation behind our study and presenting the research questions. This is followed by a brief related work part. Then, a detailed description of the survey methodology is introduced. The subsequent section presents the results and provides a thorough discussion of the findings. Finally, the paper concludes by summarizing the major outcomes of the study, highlighting their implications for integrating LLMs into academic assignment preparation, and proposing refinements to our approach to enhance its usability and acceptance.

2 Motivation and Research Questions

Following the second assignment in December 2022, we observed a noticeable improvement in the quality of some student essays compared to their first submissions. Several essays exhibited a writing style that differed from what is typically expected of computer science students. In previous years, such discrepancies pointed to the involvement of ghostwriters. The new ghostwriters are chatbots, particularly ChatGPT.

After examining the impressions our students have regarding the use of GenAI in education, we felt prepared to undertake an extensive case study related to the integration of GenAI in the preparation and delivery of paper assignments [5]. An experimental framework that allowed students to prepare their assignments with complete, partial, or no reliance on the LLM was carefully designed. Following clearly defined rules and strategies, students systematically presented their research process and submitted their final assignments [15]. This approach enabled us to evaluate not only the outcomes of each method but also students' engagement with and interaction with GenAI tools throughout the assignment lifecycle.

Since this experiment was both novel and unconventional, we aimed to explore students' perceptions of the experience and gain insight into their attitudes toward GenAI-supported academic work. The course was enrolled by 150 students, only 100 of whom were active throughout the whole semester. Aware of the challenges they faced with their first structured integration of large language models (LLMs) for paper assignment preparation, we conducted a survey to examine student viewpoints on the role of GenAI in shaping the overall integration process.

The survey addresses the following core research questions:

1. How do students perceive the integration of generative AI tools in the preparation and submission of academic paper assignments?
2. What are the key challenges and ethical considerations students associate with this integration?
3. How can educational institutions design more effective, transparent, and ethically sound approaches for incorporating GenAI into academic workflows?

3 Perception of GenAI Integration in Education

This section of the paper presents students' impressions and perceptions of the application of GenAI, and especially LLMs, in an educational context.

A 2023 analysis conducted at two universities, one in North Macedonia and the other in Sweden, gathered feedback from 112 students through a survey that explored how and why they use large language models for learning and completing writing and computer programming assignments [17]. The study addresses key concerns surrounding LLM use in education, such as copyright infringement, bias in training data, threats to human creativity, and the spread of misinformation. Based on their findings, the authors propose an original strategy for integrating LLMs into educational settings, aiming to enhance learning outcomes by promoting research, creativity, academic integrity, and critical thinking.

In their study, Chan and Hu explored university students' perceptions of GenAI technologies like ChatGPT in higher education, focusing on the benefits and challenges

of GenAI integration [18]. A survey of 399 Hong Kong students revealed a generally positive outlook on GenAI's role in personalized learning, writing, brainstorming, and research. However, students expressed concerns about accuracy, privacy, ethical implications, and the impact on personal development and careers. The study stresses the importance of understanding student perceptions, as they influence learning approaches and outcomes, and suggests informed policies that integrate GenAI into education.

Through a case study with 59 library and information science students in Zimbabwe, the research highlights growing awareness of LLMs and concerns over misinformation, bias, and ethical issues [19]. The authors emphasize the need for ethical training and AI literacy programs for both students and educators.

Conducted with 61 students in an ICT course, the qualitative study examines undergraduate students' ethical perceptions of AI tools in academic contexts, highlighting the need for robust public policies to safeguard ethics and privacy [20]. The findings emphasize the importance of teaching ethical awareness and fostering critical thinking around AI use in learning environments. The authors advocate for continued research involving both students and educators to develop more effective and ethically grounded frameworks for integrating AI into education.

Through a case study involving 13 mechanical and management engineering master students from Italy, the research evaluates the quality of essays generated with LLM assistance, the effectiveness of current detection tools, and students' perceptions of LLMs as learning aids [21]. Student feedback, gathered through a questionnaire informed by established technology acceptance models, indicated a generally positive view of LLMs in education. The study offers valuable insights and practical recommendations for educators, while also pointing to the need for further research on integrating LLMs into academic settings responsibly.

As part of the assignment, student groups analyzed a fictional ethical dilemma using both in-class discussions and an LLM of their choice as a discussion partner [22]. Students generally found the LLM helpful for expanding their understanding and exploring diverse perspectives, though they noted its limitations, such as offering impractical suggestions and refraining from making decisions. Importantly, the exercise required students to take responsibility for choosing a course of action, reinforcing critical thinking.

All these case studies raise several ethical concerns that need to be addressed. They include issues of copyright infringement, bias in training data, and the potential threat to human creativity and personal development, as overreliance on LLMs could hinder independent thinking. Additionally, there are concerns about the spread of misinformation, accuracy, and privacy, especially regarding how LLMs handle personal data.

The ethical use of these tools requires fostering critical thinking and ethical awareness among students, ensuring they understand the limitations of LLMs, such as their inability to make definitive decisions or provide fully reliable content. Furthermore, there is a growing need for comprehensive AI literacy programs and policies to guide the responsible and effective integration of LLMs in academic environments.

4 Methodology

This paper presents students' perceptions regarding the approach to defining the rules for using GenAI in the preparation of paper assignments, the strategy of essay submission integrated within the quizzes on the Moodle Learning Platform, as well as the educational application of GenAI. It begins with an overview of several case studies on the application of GenAI for educational purposes, with a particular focus on student perceptions regarding how the implementation was carried out.

After presenting the related research and summarizing the findings, the paper briefly introduces our original experiment, highlighting the key impressions of teachers on the overall approach. Our study distinguishes itself by its focus on the direct integration of GenAI into the academic workflow, specifically the preparation and submission of paper assignments.

While much of the existing literature on GenAI's educational applications concentrates on its potential for assisting in research, content generation, or enhancing creative tasks, our research explores the operationalization of GenAI tools within the formal academic structure, examining how students engage with GenAI as a part of their assignment process and its implications on their learning experience. This original angle allows for a deeper understanding of the practical challenges students face when interacting with GenAI and provides a nuanced perspective on how such tools can be better integrated into academic curricula.

Following this, we present the survey, which was designed to gather student feedback on our approach. The survey is self-administered and collects primary data [12]. Its main purpose is to identify the key obstacles students faced during the structured use of GenAI-facilitated preparation of their written assignments and to investigate their perception about ethical use of GenAI in their academic development [9]. The survey followed a detailed analysis of the major mistakes students have made since the introduction of the approach. It consists of 25 questions, divided into six clusters:

- Time and productivity: These questions estimate the time required to conduct research and the number of references reviewed.
- Use of GenAI in the work process: This cluster includes two diagnostic questions about students' use of large language models (LLMs) in assignment preparation.
- Moodle quizzes for demonstrating GenAI integration and their instructions: These questions assess the applicability of the quizzes used to demonstrate research methods through GenAI.
- Essay submission process: These five questions evaluate whether students understood how to upload their essays.
- Problems and suggestions for improvement: After evaluating the quizzes, this section invites students to provide feedback on the challenges they faced and suggestions for enhancing the approach.
- Perception of GenAI in education: The final questions replicate an initial poll to gauge students' overall attitudes toward the use of LLMs in their studies.

5 Results and Discussions

The survey aimed to assess both the ease of use and the effectiveness of GenAI tools for academic tasks, as well as to explore students' broader attitudes toward AI-assisted education. The responses to all questions were text-based, with most questions offering predefined options that were arranged in ascending order, from the least favorable to the most supportive and ethical. Additionally, the open-ended questions in the “Problems and suggestions for improvement” section provided students with an opportunity to propose ways to better integrate GenAI into project assignment preparation.

The survey was completed by 43 students, of whom 62.79% expressed satisfaction with the proposed approach of integrating GenAI into the course. However, 9.30% believed that there was no need to demonstrate research and communication with the LLMs, arguing that the result was more important than the process. This perspective may stem from their absence during lectures, which likely led to a lack of understanding of the purpose behind the quizzes and the structured approach to text preparation.

5.1 Quantitative Analysis of the Survey

Based on the students' responses, several interesting conclusions can be drawn. First, there appears to be no strong correlation between the answers, as measured by Cramér's V—a statistic that assesses the association between two variables (see Fig. 1). This lack of high correlation is actually a positive outcome when the survey is designed to assess different, independent topics. It suggests that each question captures distinct information, without significant overlap or redundancy. As a result, the survey reflects a broader range of competencies and perspectives, enhancing both the depth and reliability of the analysis. In short, the low correlation between responses indicates that each question adds unique value to the overall understanding.

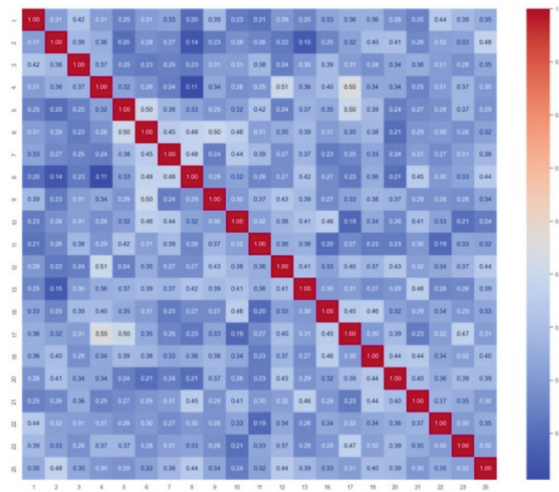


Fig. 1. Cramér's V correlation between single-choice questions.

It is also worth noting that the responses to the questions are generally well distributed (see Fig. 2), which is a positive sign. This suggests that the survey encouraged a range of answers and did not steer participants toward any particular response.

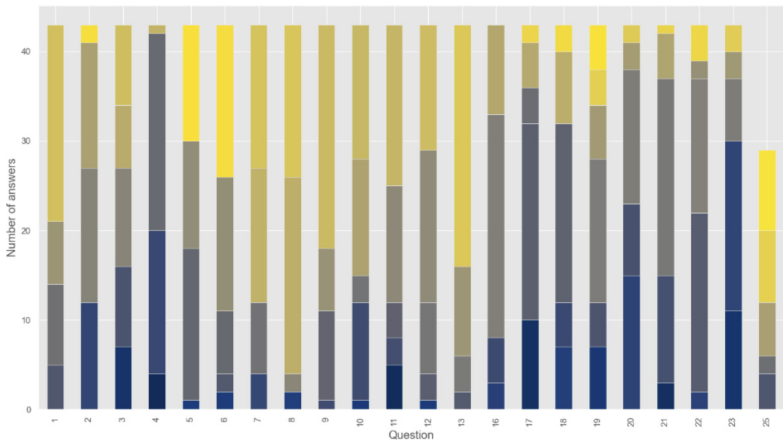


Fig. 2. Stacked Bar Plot: Number of answers per single-choice questions.

Most students needed only a few hours to prepare their reports, with more than 60% completing the task in under two hours. This suggests that the workload was manageable and that the students were able to efficiently organize and present their findings within a relatively short period of time.

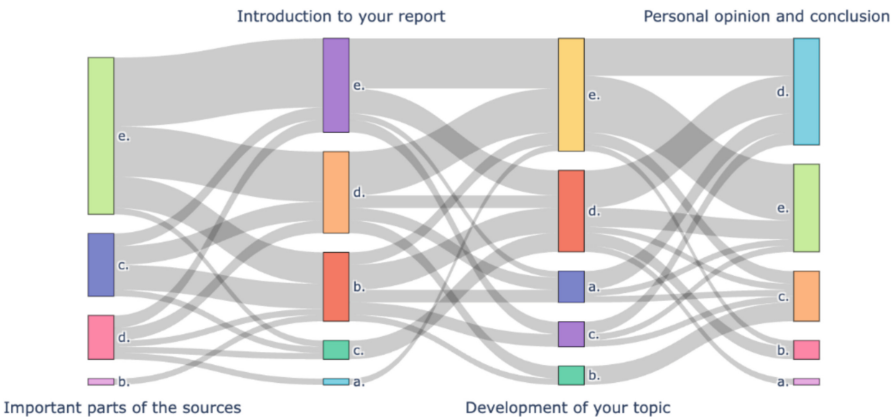


Fig. 3. Sankey Diagram.

In response to the questions about how they completed the quiz, most students reported the beginning with thorough research from reliable sources, selectively incorporating only the most relevant information (see Fig. 3). They described a structured

approach: explaining the topic they intended to cover, outlining their report's content, elaborating on the subject matter, and citing sources appropriately. Many also summarized the key insights they gained and included their personal reflections on the topic. These responses indicate a thoughtful, methodical approach that reflects both academic integrity and genuine engagement with the material.

Regarding the use of large language models (LLMs) during their preparation, most students used these tools primarily to gather initial information about their assigned topics. They then conducted independent research, integrating insights from AI with their own findings. Many emphasized the importance of verifying AI-generated content, especially since the quiz required citations from credible sources. Students demonstrated a critical approach, aware of the potential inaccuracies—or “hallucinations”—associated with AI tools.

While overall satisfaction with the process was high, students suggested that future improvements could focus on enhancing clarity, simplifying instructions, and promoting independent thinking. Most students used ChatGPT for their project work, citing its accessibility, familiarity, and effectiveness. A few also mentioned using tools like Gemini, Copilot, DeepSeek, Claude, and Bard for additional support. Some chose not to use LLMs at all, preferring to maintain full authenticity in their work.

A significant majority (72.09%) recognized the challenges of ghostwriting and plagiarism when GenAI is heavily relied upon for assignments. Only one student reported not validating AI-generated answers, while 28.57% cited hallucinations as a major concern. Although 9.30% were hesitant to rely on LLMs, over 40% used them to generate initial ideas, and more than half collaborated with AI throughout their work. In terms of overall attitudes toward GenAI in education, approximately 20% used it for personalized learning or content creation, about one-quarter for research or programming assistance, and the majority saw it as a valuable tool for tutoring and training.

5.2 Sentiment Analysis of Open Questions

The sentiment analysis was performed according to the answers to the following three open-ended questions:

1. Please describe the problems you encountered during the implementation of the project task.
2. What do you suggest will facilitate the completion of the assignment that integrates Large Language Models (LLMs)?
3. Which LLMs do you use in your everyday work and which ones did you use in your research and why?

Sentiment Analysis of Encountered Challenges

The first question was answered by 36 students. The majority of 41.67% of them had a positive sentiment expressing no issues and a smooth experience. Their answers were almost identical, such as: “I didn’t encounter any problems during the implementation of the project task.”, “Honestly, I didn’t encounter any problems.” and “Everything was in perfect order.”. They expressed clear understanding, no technical issues, or felt confident in performing the task.

Neutral responses were given by 30.56% of all participants. These comments generally addressed minor issues, offered constructive feedback, or shared neutral observations. Examples include: "I didn't encounter specific problems, but my main dilemma was how much of the project to do myself versus how much to delegate to the LLM," and "No major issues overall—different topics require different levels of effort. One suggestion: make the quiz instructions clearer so all students understand how to complete it." Another noted, "There were too many requirements, but I wouldn't say there were problems." One comment, unrelated to LLMs but important, highlighted content access restrictions: "Some questions couldn't be answered through search, so we used approximate content. Google also censors some results, so deeper searches across multiple engines were needed"—a clear ethical concern.

The remaining 27.78% expressed negative sentiment, focusing mainly on unclear quiz instructions, difficulty finding relevant content, and time-consuming or repetitive tasks. Some also mentioned challenges with academic writing and citation, especially for first-time experiences. Student remarks included: "Instructions were unclear, and tasks were too long and repetitive," "I had to revisit the task several times to understand what was expected," and "Working with an LLM is a dynamic process. It's hard to define fixed questions and answers since the interaction involves constant input to guide the model's output."

Sentiment Analysis with Suggestions for Improvement

The second question was answered by 37 students. A majority (54.05%) expressed neutral sentiment, often responding with "No idea" or "No comment." However, some responses stood out, such as: "More attention should be given to verifying the reliability of information—ask critical questions; we shouldn't let our brains 'sit idle,'" and "Rather than simplifying, it would be interesting to form groups, each using a different LLM on the same topic, then compare the results." The first comment raises ethical awareness, while the second offers an analytical and exploratory suggestion with academic potential.

Preserving the current approach was supported by 37.84% of students. Encouraging feedback included: "Nothing needs to be simplified—minimal effort yields significant knowledge," and "The current approach is good; it helped me with other papers too." These responses highlight that the existing structure supports meaningful learning, fosters independence, and builds transferable skills.

Only two students expressed critical views, both cautious about LLM integration. One suggested, "There should be a balance between using LLMs and meeting requirements," while the other noted, "LLMs offer fast, helpful information but pose risks like inauthentic work and potential plagiarism." These responses reflect a mature understanding of the ethical trade-offs, acknowledging both the benefits and the concerns that may impact academic integrity.

Summary of Student Reflections About Usage Patterns and Preferences for LLMs

The last question does not reflect any sentiment. Instead, the answers were summarized quantitatively and qualitatively.

The most notable finding is the overwhelming preference for ChatGPT, which was used or mentioned by nearly all respondents (23 out of 27). It is widely seen as an accessible and effective tool for both academic and everyday tasks, including content generation, summarization, answering questions, and explaining technical concepts like

code. DeepSeek was the second most used, especially valued for in-depth explanations. Microsoft Copilot and GitHub Copilot were primarily referenced for coding tasks, highlighting their role in technical work. Claude and Gemini were mentioned less frequently but were used for specific purposes, such as research support or clearer code explanations.

Most respondents used ChatGPT for research—generating ideas, gathering quick information, and structuring arguments—regarding it as a reliable aid. Other LLMs like DeepSeek, Claude, and Gemini were used by a smaller group, often in combination with ChatGPT depending on the task.

In terms of overall utility and preference, ChatGPT stood out due to its accessibility and familiarity. Respondents used it for both academic and practical purposes, such as writing reports, creating content, and even drafting emails. Though used less often, Claude and Gemini were appreciated for tackling specific challenges like refining content or generating technical explanations.

A few respondents reported minimal or no use of LLMs, with some choosing to maintain the authenticity of their work and avoid standardized outputs. Interestingly, several students indicated that combining different LLMs worked best, using tools like ChatGPT alongside Claude or DeepSeek to address different aspects of their tasks. This shows a thoughtful, task-based approach and an understanding of each model's strengths.

6 Conclusions and Future Work

The integration of GenAI within the Computer Ethics course followed a very systematized and specific approach, which stimulated students to demonstrate how they did the research and prepared the written assignments [8]. The survey done to assess their attitudes towards GenAI indicate a generally positive reception among students toward the integration of chatbot-based tools, particularly for their efficiency, accessibility, and potential to support independent learning. Almost all reviewed students used LLMs for initial idea generation and report writing, while some employed a combination of LLMs for task-specific needs, demonstrating adaptability and critical engagement. This reflects a level of critical engagement and adaptability.

While a minority expressed concerns about the authenticity of AI-assisted work and questioned the need for structured interaction with GenAI, the overall sentiment emphasized the educational value and practical utility of these tools. Although a minority expressed skepticism about the need for structured interaction with LLMs or raised concerns over authenticity and academic integrity, the overall sentiment reflected appreciation for the educational value and practical utility of these tools. At the same time, the survey also highlights areas for improvement, such as clearer instructions, better guidance on responsible AI use, and fostering critical evaluation skills.

These insights affirm the potential of GenAI to enhance learning when integrated thoughtfully and ethically into coursework. The survey also revealed key areas for improvement, including the need for clearer instructions, stronger guidance on ethical AI use, and support for developing critical evaluation skills. These findings confirm that GenAI can enhance student learning when integrated thoughtfully, with attention to transparency and academic integrity.

Future work will focus on refining the instructional framework based on student feedback. Planned modifications include updated guidelines for GenAI use in assignments,

more detailed ethical use instructions, and improved feedback mechanisms within the Moodle platform. Long-term, the research will expand to examine disciplinary differences in student engagement with GenAI, as well as the sustained impact of AI-assisted learning on academic performance, autonomy, and ethical reasoning. This iterative and student-centered approach aims to evolve best practices for the ethical and pedagogically meaningful integration of GenAI into higher education.

Acknowledgement. This work was supported in part by grants from the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje.

References

1. Sengar, S.S., et al.: Generative artificial intelligence: a systematic review and applications. *Multimedia Tools and Applications*, 1–40 (2024)
2. Sedkaoui, S., Benaichouba, R.: Generative AI as a transformative force for innovation: a review of opportunities, applications and challenges. *European J. Innov. Manage.* (2024)
3. Zhai, X.: Transforming teachers' roles and agencies in the era of generative AI: Perceptions, acceptance, knowledge, and practices. *J. Sci. Educ. Technol.* 1–11. (2024)
4. Shailendra, S., Kadel, R., Sharma, A.: Framework for adoption of generative artificial intelligence (GenAI) in education. *IEEE Transactions on Education* (2024)
5. Zdravkova, K., Ilijoski, B.: Preventing academic dishonesty originating from large language models. In: *Advances in ICT Research in the Balkans. BCI 2024. Communications in Computer and Information Science*, vol 2391. Springer, Cham (2025)
6. Kasneci, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Diff.* **103**, 102274 (2023)
7. Yan, L., et al.: Practical and ethical challenges of large language models in education: a systematic scoping review. *Br. J. Edu. Technol.* **55**(1), 90–112 (2024)
8. Zdravkova, K., Ilijoski, B.: The impact of large language models on computer science student writing. *Int. J. Educ. Technol. High. Educ.* **22**(1), 1–24 (2025)
9. Schneider, J.: Explainable generative ai (genxai): A survey, conceptualization, and research agenda., *Artif. Intel. Rev.* **57**(11), 289 (2024)
10. Luo, J.: How does GenAI affect trust in teacher-student relationships? Insights from students' assessment experiences. *Teaching in Higher Education*, 1–16 (2024)
11. Bozkurt, A.: GenAI et al. Cocreation, authorship, ownership, academic ethics and integrity in a time of generative AI. *Open Praxis* **16**(1), 1–10 (2024)
12. Bispo, E.L., dos Santos, S.C., De Matos, M.V.: Equity issues derived from use of large language models in education. In: *International Conference on New Media Pedagogy*, pp. 425–440. Springer Nature Switzerland, Cham (2023)
13. Ferrara, E.: GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *J. Computat. Soc. Sci.* **7**(1), 549–569 (2024)
14. Abdaljalael, M., et al.: A multinational study on the factors influencing university students' attitudes and usage of ChatGPT. *Scientific Reports* **14**(1), 1983 (2024)
15. Zdravkova, K., Dalipi, F., Ahlgren, F., Ilijoski, B., Olsson, T.: Unveiling the impact of large language models on student learning: a comprehensive case study. In: *2024 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1–8. IEEE (2024)
16. Hellas, A., Leinonen, J., Leppänen L.: Experiences from integrating large language model chatbots into the classroom. In: *Proceedings of the 2024 ACM Virtual Global Computing Education Conference V. 1* (2024)

17. Zdravkova, K., Dalipi, F., Ahlgren, F.: Integration of Large language models into higher education: a perspective from learners. In: 2023 International Symposium on Computers in Education (SIIE), pp. 1–6. Setúbal, Portugal (2023)
18. Chan, C.K.Y., Hu, W.: Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *Int. J. Educ. Technol. High. Educ.* **20**(1), 43 (2023)
19. Chigwada, J., Pasipamire, N.: Perception and use of large language models by library and information science students. *Int. J. Librarians.* **9**(3), 75–89 (2024)
20. da Silva, M., et al.: Ethics and AI in higher education: A study on students' perceptions. In: International Conference on Information Technology & Systems, pp. 149–158 (2024)
21. Bernabei, M., Colabianchi, S., Falegnami, A., Costantino, F.: Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances. *Comp. Edu. Artif. Intel.* **5**, 100172 (2023)
22. Grande, V., Kiesler, N., Francisco R.M.A.: Student Perspectives on Using a Large Language Model (LLM) for an Assignment on Professional Ethics. In: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, pp. 478–484 (2024)



Comparative Analysis of Instructor and AI Assessments: Objectivity, Biases, and Impact on Academic Grading

Teresa Torres-Coronas^(✉) 

Universitat Rovira i Virgili, 43007 Tarragona, Catalonia, Spain
teresa.torres@urv.cat

Abstract. This article examines the role of artificial intelligence in academic assessment, with particular focus on its implications for grading accuracy. It presents a comparative analysis of 46 undergraduate reports, each evaluated independently by a human instructor and an AI-based system. A quantitative methodology combining correlation analysis, Bland-Altman plots, and paired t-tests was employed to compare human and AI evaluations. The results reveal discrepancies between the two sets of scores, which suggest that AI systems fail to adequately assess qualitative dimensions such as originality, critical thinking, and contextual relevance. While AI offers advantages in consistency, scalability, and efficiency, it often fails to capture the pedagogical depth and interpretive insight provided by human evaluators. In light of these findings, the article proposes a hybrid assessment framework that combines the precision of AI with the interpretive competences of human instructors.

Keywords: AI Grading · Academic Grading · Biases

1 Introduction

Artificial intelligence (AI) has become an integral force in educational environments, reshaping how academic assessments are performed. The rise of AI-based grading systems promises enhanced efficiency, consistency, and objectivity. However, as this technology becomes more prominent, critical questions arise regarding its accuracy, fairness, potential biases, and ability to evaluate student work [1–3].

While human instructors can adapt their evaluations based on student context and effort, AI systems like ChatGPT rely on statistical patterns learned from large datasets. Although they can generate contextually relevant responses, they still lack genuine understanding and are limited in their ability to interpret contextual meaning or originality in student work [4, 5]. This raises a fundamental question: To what extent can AI-driven evaluation models replicate the holistic and interpretative judgment traditionally applied by human instructors?

Although, AI assessments have the potential to rationalize grading processes and provide standardized evaluations, they often lack the contextual awareness that human

instructors possess [6, 7]. Educators can account for creativity, effort, and originality, factors that enrich students' reports. Recent discussions [8] have highlighted AI's limited capacity to assess relevance, linguistic appropriateness, and the effective use of supporting evidence, which are dimensions that require interpretative flexibility. The systematic literature review by [8, 9] indicates that AI systems lack the academic rigor required for producing scholarly writing. As a result, their ability to properly evaluate research reports remains limited. This contrasts with previous studies that highlight the instructional potential of AI systems in providing feedback to students [9, 10].

Given these concerns, understanding how to integrate its capabilities while addressing its limitations becomes essential for future academic frameworks regarding the regulation of AI usage in student evaluation processes. In response to this challenge, this study examines the differences between AI-generated assessments and those carried out by human instructors.

To explore these differences, this research applies a structured quantitative methodology. It combines statistical analyses and paired samples t-tests to measure discrepancies across key assessment dimensions such as objective alignment, relevance, language quality, and presentation. Preliminary results show that AI tends to assign significantly higher and more homogeneous scores, particularly in language and formal aspects, whereas instructors apply more accurate and context-sensitive judgments. Similar trends were also observed in [8], reinforcing the present study's findings. These findings highlight the need to balance the efficiency of automated systems with the qualitative judgment of human instructors.

The following sections present the study's methodology, key findings, and a critical discussion of the differences between AI and human evaluation, concluding with implications for hybrid assessment models.

2 Methodology

2.1 Sample Data

This research is based on data from assessments conducted by both human instructors and AI on students' academic work. As part of an academic assignment, undergraduate students conducted market research on the acceptance of technological implants aimed at enhancing human capabilities. Based on their findings, students submitted a writing report. The primary objective of this task was to assess students' academic writing skills within the framework of research-based reporting. A total of forty-six reports were independently evaluated by a faculty member and by an AI-based assessment system (ChatGPT, version 4).

To enable a direct comparison between human and AI evaluations, all reports were evaluated using a standardized rubric based on the official academic writing standards established by the university. This structured approach ensured consistency in grading, and facilitated a systematic analysis of potential discrepancies. Table 1 outlines the criteria, their corresponding weights, and definitions.

Table 1. Evaluation criteria, weigh and definition.

Criteria	Weigh	Description
Objective	15%	The report effectively addresses the assigned topic and aligns with its intended purpose, such as evaluation, analysis, or recommendations
Relevance	10%	The ideas and information presented are clearly pertinent to the topic under discussion
Supporting elements	25%	The report includes additional material, such as bibliographic references, to substantiate the presented ideas, with appropriate citation and formatting
Structure and coherence	10%	The content follows the standard academic format (introduction, methodology, results, and conclusions) and maintains a logical and coherent sequence throughout
Language and style	30%	The writing is appropriate for an academic audience, using accurate technical language and demonstrating grammatical and typographical correctness
Presentation	10%	The report follows specific format requirements, such as word count limits, and it is professionally structured for clarity and readability

The AI system (ChatGPT, version 4) was prompted with the following instruction: *Evaluate this university research paper using the provided rubric, considering the description of each criterion and its respective weight, and assigning a final score on a 10-point scale.* This prompt was designed to operationalize the rubric in a manner consistent with the assessment standards used by human evaluators. To support a context-aware evaluation, the AI was also provided with the original assignment guidelines and learning objectives given to students.

This approach aimed to simulate a realistic and replicable AI evaluation scenario, aligning the model's outputs with the structured expectations of academic grading. While generative models lack interpretative reasoning, previous research [11] suggests that systematic discrepancies between large language model, such as ChatGPT, and human grading can be mitigated through prompt design. Providing rubric-based instructions serves not only to constrain the model's output but also to align its scoring behaviour with human evaluation standards.

2.2 Statistical Analysis

A quantitative methodology was applied to assess grading discrepancies between instructor and AI evaluations, analyse correlations among assessment criteria, and quantify the contribution of specific evaluation dimensions to overall score divergence (see Table 1). This approach allowed for a systematic and objective examination of potential biases in AI-driven grading compared to human evaluation. The results will contribute to a

clearer understanding of the reliability, consistency, and limitations of AI-based assessment systems in higher education. These findings are consistent with previous research that has questioned the effectiveness of tools like Gemini and ChatGPT when used as academic writing assistants or evaluators [8, 11].

Relationship Between AI and Instructor Grades

The initial comparison between AI-generated and instructor-assigned scores revealed a Pearson correlation coefficient of $r = 0.19$, a low value also observed in previous research [12]. This suggests that while there may be some general trend between the grading systems, linear correlation alone is insufficient to establish agreement.

For that reason, a Bland-Altman analysis was conducted to assess the level of agreement and detect potential bias between the two scoring methods. Bland-Altman analysis enables a visual and quantitative assessment of the consistency and direction of grading discrepancies, a key aspect when comparing evaluation methods [8].

Figure 1 illustrates the difference between instructor and AI scores plotted against the mean of each pair. Instructor scores are consistently lower, with a mean difference of approximately -2.0 .

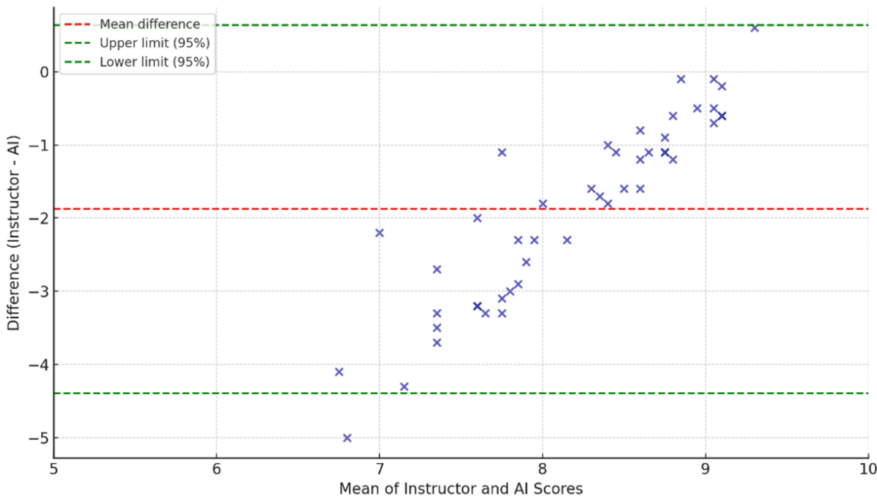


Fig. 1. Bland-Altman plot: Instructor vs AI.

Most data points fall below the zero-difference line, with 95% limits of agreement ranging between -3 and -1 , indicating a scoring bias in favour of the AI. This supports prior findings that AI systems often assign higher scores and perform less consistently in tasks requiring interpretative analysis, such as written or project-based assessments, compared to more objective assessments formats [13].

In contrast, human evaluations show greater variability, likely reflecting a more context-sensitive interpretation of student performance, influenced by subjective and

individual factors [3]. Overall, the Bland-Altman reveals consistent divergence and motivates a criterion-level analysis of the grading discrepancies.

Criterion-Based Correlation Analysis

To better understand the nature of the grading discrepancies, a correlation analysis was performed on the individual evaluation criteria (Table 2). This analysis helps to determine whether differences between AI and instructor assessments are linked to specific aspects of the grading rubric. Additionally, the internal consistency of both the instructor's and the AI's scores were examined to understand how each one prioritizes key components of academic quality.

Table 2. Pearson correlations between instructor and AI evaluation criteria

		Correlations											
		ObjIns	ObjIA	RelevIns	RelevIA	SupplIns	SupplIA	StrucIns	StrucIA	LangIns	LangIA	PresIns	PresIA
ObjIns	Pearson Correlation	1.000	-.009	.623	NaN	-.032	-.204	.232	-.104	.049	-.148	-.261	.232
	Sig. (2-tailed)		.951	.000	NaN	.832	.174	.122	.497	.749	.326	.080	.121
	N	46	46	46	46	46	46	46	45	46	46	46	46
ObjIA	Pearson Correlation	-.009	1.000	-.021	NaN	.046	.281	.069	.229	.296	.442	.020	.236
	Sig. (2-tailed)	.951		.890	NaN	.759	.059	.648	.131	.046	.002	.893	.115
	N	46	46	46	46	46	46	46	45	46	46	46	46
RelevIns	Pearson Correlation	.623	-.021	1.000	NaN	-.210	-.193	.183	.038	.065	-.150	.069	-.115
	Sig. (2-tailed)	.000	.890		NaN	.162	.198	.224	.802	.667	.319	.649	.446
	N	46	46	46	46	46	46	46	46	45	46	46	46
RelevIA	Pearson Correlation	NaN	NaN	NaN	NaN	-.Infinity	-.Infinity	-.Infinity	+.Infinity	-.Infinity	-.Infinity	+.Infinity	+.Infinity
	Sig. (2-tailed)	NaN	NaN	NaN	NaN	.000	.000	.000	.000	.000	.000	.000	.000
	N	46	46	46	46	46	46	46	45	46	46	46	46
SupplIns	Pearson Correlation	-.032	.046	-.210	-.Infinity	1.000	.164	-.059	.000	.043	.087	-.024	-.003
	Sig. (2-tailed)	.832	.759	.162	.000		.277	.695	.998	.777	.565	.872	.986
	N	46	46	46	46	46	46	46	45	46	46	46	46
SupplIA	Pearson Correlation	-.204	.281	-.193	-.Infinity	.164	1.000	.011	-.041	.299	.636	.347	.328
	Sig. (2-tailed)	.174	.059	.198	.000	.277		.943	.791	.043	.000	.018	.026
	N	46	46	46	46	46	46	46	45	46	46	46	46
StrucIns	Pearson Correlation	.232	.069	.183	-.Infinity	-.059	.011	1.000	-.130	.201	.216	.312	-.159
	Sig. (2-tailed)	.122	.648	.224	.000	.695	.943		.395	.181	.150	.035	.293
	N	46	46	46	46	46	46	46	45	46	46	46	46
StrucIA	Pearson Correlation	-.104	.229	.038	+.Infinity	.000	.041	-.130	1.000	.205	-.041	.143	-.058
	Sig. (2-tailed)	.497	.131	.802	.000	.998	.791	.395		.176	.791	.349	.706
	N	45	45	45	45	45	45	45	45	45	45	45	45
LangIns	Pearson Correlation	.049	.296	.065	-.Infinity	.043	.299	.201	.205	1.000	.275	.417	.305
	Sig. (2-tailed)	.749	.046	.667	.000	.777	.043	.181	.176		.064	.004	.040
	N	46	46	46	46	46	46	46	45	46	46	46	46
LangIA	Pearson Correlation	-.148	.442	-.150	-.Infinity	.087	.636	.216	-.041	.275	1.000	.346	.172
	Sig. (2-tailed)	.326	.002	.319	.000	.565	.000	.150	.791	.064		.018	.254
	N	46	46	46	46	46	46	46	45	46	46	46	46
PresIns	Pearson Correlation	-.261	.020	.069	+.Infinity	-.024	.347	.312	.143	.417	.346	1.000	-.003
	Sig. (2-tailed)	.080	.893	.649	.000	.872	.018	.035	.349	.004	.018		.986
	N	46	46	46	46	46	46	46	45	46	46	46	46
PresIA	Pearson Correlation	.232	.236	-.115	+.Infinity	-.003	.328	-.159	-.058	.305	.172	-.003	1.000
	Sig. (2-tailed)	.121	.115	.446	.000	.986	.026	.293	.706	.040	.254	.986	
	N	46	46	46	46	46	46	46	45	46	46	46	46

Note: Significant correlations at the 0.05 level (2-tailed) are indicated. NaN and Infinity values correspond to technical anomalies in AI assessments of relevance.

Correlations Among Instructor Evaluations

The instructor evaluations showed significant relationships between some assessment criteria. A strong positive correlation was observed between Objective and Relevance scores ($r = 0.623$, $p < 0.001$), indicating that instructors who rated the achievement of objectives highly also tended to assign higher relevance scores. This reflects internal coherence in human evaluation, where thematic alignment and fulfilment of assignment goals are considered jointly.

A moderate correlation between Structure and Presentation ($r = 0.312$, $p = 0.035$) showed that better document organization was associated with improved presentation. From the instructor's perspective structural clarity contributes to a more professional

appearance. Additionally, Language and Presentation were significantly correlated ($r = 0.417$, $p = 0.004$), reinforcing the idea that linguistic accuracy plays a key role in shaping the evaluator's perception of formal quality. In contrast, the remaining correlations among instructor-assigned criteria were weak or non-significant. This result points out a more independent evaluative approach across aspects such as supporting evidence and language use when considered individually.

Correlations Among AI Evaluations

AI-based evaluations demonstrated strong internal consistency across technical aspects of the assignments. A high correlation was observed between Language and Supporting Elements ($r = 0.636$, $p < 0.001$), which suggests that the system consistently linked linguistic quality with the inclusion and accuracy of bibliographic references. This outcome is aligned with a more procedural and rule-driven logic. Language was also positively associated with Objective scores ($r = 0.442$, $p = 0.002$), a finding that indicates that well-written texts were more likely to be judged by the AI as successfully addressing the task.

Similarly, a significant relationship between Language and Presentation ($r = 0.346$, $p = 0.018$) reflects the model's tendency to equate formal linguistic correctness with document quality. Supporting Elements showed moderate correlations with Presentation ($r = 0.328$, $p = 0.026$) and Language ($r = 0.299$, $p = 0.043$), underscoring a consistent interdependence among citation practices, language precision, and formatting.

In contrast to human evaluations, which display greater variability and contextual sensitivity, AI assessments appear to rely more heavily on surface-level formal criteria. This behaviour reflects the automated system's strict adherence to rubric parameters and aligns with prior studies that point to the limitations of AI in capturing academic rigor, domain expertise, conceptual synthesis, and human-like reasoning [6].

Correlations Between Instructor and AI Evaluations

Cross-evaluator comparisons revealed limited alignment between instructor and AI scoring. A low positive correlation was found between Language scores assigned by the instructor and those assigned by the AI ($r = 0.296$, $p = 0.046$), which reflects some degree of overlap in the assessment of linguistic quality.

Similarly, Presentation scores showed a moderate correlation ($r = 0.346$, $p = 0.018$), suggesting that both evaluators tended to identify similar formal features related to document formatting and clarity. In addition, instructor-assessed Language scores moderately correlated with the AI's Presentation ratings ($r = 0.305$, $p = 0.040$), a result that highlights a shared sensitivity to the relationship between linguistic precision and perceived professionalism.

Beyond these cases, most correlations between human and AI scores were weak or non-significant, showing substantial differences in evaluative frameworks and grading logic. It is also worth noting that the AI's evaluation of Relevance produced undefined or infinite correlation coefficients, likely due to a lack of score variation, a limitation that compromised the statistical reliability of that criterion.

Paired Samples t-Test: Instructor vs AI Scoring Differences

To quantify the magnitude and statistical significance of the discrepancies, a paired

samples t-test was conducted for each evaluation criterion (Table 3). This test assessed whether the mean differences were statistically significant.

Table 3. Paired Samples Test

Paired Samples Test									
		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	S.E. Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	ObjIns - ObjIA	-.22	.53	.08	-.38	-.06	-2.79	45	.008
Pair 2	RelevIns - RelevIA	-.12	.24	.04	-.19	-.05	-3.41	45	.001
Pair 3	SuppIns - SuppIA	-.43	.95	.14	-.71	-.15	-3.07	45	.004
Pair 4	StrucIns - StrucIA	-.05	.27	.04	-.13	.04	-1.15	44	.256
Pair 5	LangIns - LangIA	-.96	.37	.06	-1.07	-.85	-17.40	45	.000
Pair 6	PresIns - PresIA	-.10	.22	.03	-.16	-.03	-2.99	45	.004

The results show that, for five out of the six evaluated criteria, the differences between instructor and AI scores were statistically significant ($p < 0.05$). In the case of the Objective criterion, instructors tended to be slightly more stringent than the AI ($t = -2.793$, $p = 0.008$). A similar pattern was evident for Relevance ($t = -3.410$, $p = 0.001$). Supporting Elements also showed a discrepancy ($t = -3.073$, $p = 0.004$), with instructors demanding a higher standard of bibliographic evidence and citation accuracy.

In contrast, no significant difference was found in Structure scores ($t = -1.152$, $p = 0.256$), indicating relative alignment between human and AI evaluations in terms of document organization.

The most pronounced divergence emerged in the assessment of Language and Style ($t = -17.404$, $p < 0.001$), where the AI consistently awarded substantially higher scores than the instructor. Finally, for Presentation, instructors applied a more rigorous standard ($t = -2.994$, $p = 0.004$). Overall, these findings confirm that discrepancies between AI and instructor evaluations are not random, particularly in dimensions involving linguistic quality, perceived relevance, and the proper use of supporting material.

3 Discussion

The low Pearson correlation between AI-generated and instructor-assigned evaluations suggests limited alignment, though it does not fully explain differences in evaluative logic. Criterion-level analyses show that instructors emphasize contextualization, analytical depth, and coherence, while AI systems focus primarily on surface-level correctness. As noted by [11], this reflects the limited capacity of generative AI models such as ChatGPT to assess higher-order cognitive tasks. While earlier studies reported promising results regarding their effectiveness in academic assessment [14–16], the present findings raise concerns about their consistency and conceptual depth.

This divergence is further illustrated by the Bland-Altman analysis, which reveals a systematic tendency of AI to assign higher scores. The pedagogical implications are substantial: students whose work demonstrates formal correctness but lacks deeper conceptual engagement may receive inflated scores from AI systems. This could misrepresent actual learning outcomes.

Correlation patterns reinforce these concerns. Instructors showed strong associations between conceptual criteria, such as Objective and Relevance ($r = 0.623$), and Language and Presentation ($r = 0.417$), suggesting a holistic view of academic quality. They also tend to reward originality and effort, even when accompanied by minor technical flaws. In contrast, AI evaluations emphasized procedural features, such as Language, Supporting Elements, and Presentation, revealing a narrower, though internally consistent, approach.

Such a procedural bias may benefit students who follow formal conventions but disadvantage those whose innovative ideas are expressed less fluently. The gap becomes especially evident in tasks requiring critical thinking or disciplinary insight. Moreover, AI systems apply penalties consistently, without the flexibility instructors use when balancing conceptual quality against technical accuracy. This lack of interpretive sensitivity may reinforce inequities in academic evaluation.

Overall, these findings point to a fundamental misalignment between automated scoring logic and human pedagogical judgment. Rather than inconsistency, instructors' differentiated evaluations reflect a sound, content-sensitive approach. In contrast, AI systems lack the ability to provide accurate, content- or course-specific feedback [17], and may even deliver misleading guidance [18].

4 Limitations and Further Research

While this study offers valuable insights into the discrepancies between AI and instructor-based evaluation, several limitations must be acknowledged.

First, the analysis focused on a single academic task and relied on a predefined rubric. This narrow scope may limit the generalizability of findings to other disciplines, educational levels, or assignment formats. Future research should include diverse academic domains and a wider range of assessment types (e.g., reflective essays, case analyses, multimedia projects) to determine whether the observed discrepancies persist across different contexts and cognitive demands.

Second, the evaluated AI system (ChatGPT-4) followed a rigid interpretation of the rubric and lacked adaptability to contextual or creative aspects of student work. This limitation reinforces existing concerns about the inability of AI to replicate the depth of human judgment [8].

Third, methodological anomalies, particularly the undefined or infinite correlation coefficients observed in the Relevance criterion, expose vulnerabilities in current AI-based scoring models. These issues suggest a need for more robust algorithms capable of handling rubric categories that may appear ambiguous to non-human evaluators. Future studies could employ complementary techniques, such as interquartile range filtering and robust correlation metrics, to better interpret inconsistent scoring behaviour.

Additionally, this study relied on a single evaluation per submission. Alternative approaches, such as 10-shot evaluation [11], where the model assesses the same text multiple times to test internal consistency, could provide deeper insights into the stability of AI-generated scores.

Finally, the relatively small sample size and focus on a single academic field limit the scope of the conclusions. Future research should expand participant diversity, incorporate different course contexts and instructor profiles, and include inter-rater reliability measures for human evaluations to strengthen internal validity.

Looking ahead, further work should explore the development of AI training strategies capable of integrating contextual and qualitative evaluation dimensions. Longitudinal studies could also assess how AI-mediated feedback affects student learning outcomes and perceptions over time. As students increasingly use generative AI tools to support their writing [4], instructors will need new competencies to accurately assess AI-assisted work, particularly as such systems may generate plausible yet inaccurate responses [19].

In summary, the evolving relationship between AI-supported learning and assessment demands sustained empirical investigation to ensure that evaluative practices remain pedagogically meaningful, equitable, and aligned with educational values.

5 Towards a Hybrid Assessment Model

In light of these findings and limitations, a hybrid assessment model emerges as a viable and pedagogically sound alternative. In this approach, AI systems would be responsible for executing low-level, repetitive tasks—such as checking formatting, validating citations, or flagging language errors—while human instructors would retain responsibility for evaluating conceptual depth, argumentation, creativity, and contextual alignment.

This division of labour aligns with Adam Smith's principle of specialization, enhancing overall efficiency without compromising evaluative integrity. More importantly, it allows instructors to focus their expertise on areas where human judgment is indispensable and where generative AI still performs inadequately [21, 23].

Effective implementation of such hybrid models requires more than just technological integration. As noted by [22, 24], educators will need targeted professional development to adapt their assessment practices and make meaningful use of AI assistance. Institutions must also ensure that assessment frameworks remain aligned with principles of meaningful learning [20] and foster synergy between human expertise and algorithmic support [22].

By balancing automation and pedagogical sensitivity, hybrid assessment models offer a promising path forward: one that upholds academic standards while harnessing the practical benefits of AI technologies.

6 Conclusions

This study systematically compared academic evaluations generated by an artificial intelligence system (ChatGPT, version 4) with those assigned by human instructors in the context of undergraduate research reports. Through statistical analyses, including Pearson correlations, paired samples t-tests, and Bland-Altman plots, the findings revealed systematic discrepancies, particularly in assessment criteria requiring interpretative judgment.

The results indicate that, while AI systems exhibit high internal consistency in evaluating formal and technical dimensions, they consistently assign higher and more homogeneous scores than human evaluators. This suggests a lack of contextual and conceptual sensitivity in AI-generated assessments. In contrast, human instructors tend to better evaluate factors like originality, critical thinking, and contextual appropriateness, which remain beyond the interpretive capabilities of current AI models.

These observations align with [8], who argues that AI performs better in structured or objective test formats, but lacks the flexibility and cognitive depth needed for open-ended or qualitative evaluations. To address these limitations, the study proposes a hybrid assessment framework grounded in Adam Smith's principle of the division of labour, whereby repetitive and rule-based tasks, such as formatting, grammar, and citation checks, are delegated to AI systems, while tasks requiring interpretive judgment, conceptual analysis, and evaluative flexibility remain under human oversight.

This structure seeks to balance the efficiency and consistency of AI with the depth and pedagogical insight of human evaluation. There is a need for adapted assessment methods [22]. However, the implementation of such systems must be approached with caution. AI should be seen as an augmentative tool rather than a replacement for human judgment, particularly in high-stakes educational settings.

Acknowledgments. This research was funded by the Jaume Vicens Vives Distinction awarded by the Generalitat de Catalunya, which supports initiatives aimed at enhancing the quality of university teaching.

Ethical Considerations. All student submissions included in this research were used with prior informed consent and were fully anonymized to ensure compliance with institutional ethical standards and data protection regulations.

References

1. Bulut, O., et al.: The rise of artificial intelligence in educational measurement: opportunities and ethical challenges. *Computers & Society* (2024). <https://doi.org/10.48550/arXiv.2406.18900>
2. Gnanaprakasam, J., Lourdasamy, R.: The Role of AI in automating grading: Enhancing feedback and efficiency. In: *Artificial Intelligence*. IntechOpen (2024). <https://doi.org/10.5772/intechopen.1005025>
3. Yu, L., and Yu, Z.: Qualitative and quantitative analyses of artificial intelligence ethics in education using Vosviewer and Citnetexplorer. *Frontiers in Psychology* **14** (2023). <https://doi.org/10.3389/fpsyg.2023.1061778>
4. Kim, J., Yu, S., Detrick, R., Li, N.: Exploring students' perspectives on Generative AI-assisted academic writing. *Educ. Info. Technol.* **30**, 1265–1300 (2025). <https://doi.org/10.1007/s10639-024-12878-7>
5. Gruenhagen, J.H., Sinclair, P.M., Carroll, J.-A., Baker, P.R.A., Wilson, A., Demant, D.: The rapid rise of generative AI and its implications for academic integrity: students' perceptions and use of chatbots for assistance with assessments. *Comp. Educ. Artif. Intel.* **7**, 100273 (2024). <https://doi.org/10.1016/j.caeai.2024.100273>
6. Pang, W., Wei, Z.: Shaping the future of higher education: a technology usage study on generative AI innovations. *Information* **16**(2), 95 (2024). <https://doi.org/10.3390/info16020095>
7. Zhao, J., Chapman, E.I., Sabet, P.G.P.: Generative AI and educational assessments: a systematic review. *Education Research and Perspectives*. *An International Journal* **51**, 124–155 (2024). <https://doi.org/10.70953/ERPv51.2412006>
8. Coskun, T.K., Alper, A.: Evaluating the evaluators: a comparative study of AI and teacher assessments in higher education. *J. Educ. Technol.* **45**(2), 123–145 (2023). <https://doi.org/10.1344/der.2024.45.124-140>

9. Mulyanah, E.Y., Hidayat, S., Yuhana, Y., Ishak: Enhancing the academic writing process using artificial intelligence: A Bibliometric. *Humanit. Lang. Int. J. Linguist. Humanit. Edu.* **1**(5), 299–307 (2024). <https://doi.org/10.32734/djd9kz89>
10. Celik, I., Dindar, M., Muukkonen, H., Järvelä, S.: The promises and challenges of artificial intelligence for teachers: a systematic review of research. *TechTrends* **66**(4), 616–630 (2022). <https://doi.org/10.1007/s11528-022-00715-y>
11. Jauhainen, J.S., Garagorry Guerra, A.: Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovat. Edu. Teach. Int.* 1–18 (2024). <https://doi.org/10.1080/14703297.2024.2422337>
12. Kooli, C., Yusuf, N.: Transforming educational assessment: insights into the use of ChatGPT and large language models in grading. *Int. J. Human-Comp. Interact.* **41**(5), 3388–3399 (2024). <https://doi.org/10.1080/10447318.2024.2338330>
13. Tubino, L., Adachi, C.: Developing feedback literacy capabilities through an ai automated feedback tool. *ASCILITE 2022 Conference Proceedings: Reconnecting relationships through technology*, p. e22039 (2022). <https://doi.org/10.14742/apubs.2022.39>
14. Harry, A., Sayudin, S.: Role of AI in education. *Interdisciplinary Journal of Interdisciplinary Studies (Injury)*, **2**(3), 260–268 (2023). <https://doi.org/10.58631/injury.v2i3.52>
15. Mahligawati, F., Allanas, E., Butarbutar, M.H., Nordin, N.A.N.: Artificial intelligence in physics education: a comprehensive literature review. *J. Phys. Conf. Ser.* **2596**(1), 012080 (2023). <https://doi.org/10.1088/1742-6596/2596/1/012080>
16. Vittorini, P., Menini, S., Tonelli, S.: An AI-based system for formative and summative assessment in data science courses. *Int. J. Artif. Intell. Educ.* **31**(2), 159–185 (2021). <https://doi.org/10.1007/s40593-020-00230-2>
17. Ali, K., Barhom, N., Tamimi, F., Duggal, M.: ChatGPT—A double-edged sword for healthcare education? Implications for assessments of dental students. *Eur. J. Dent. Educ.* **28**(1), 206–211 (2024). <https://doi.org/10.1111/eje.12937>
18. Morjaria, L., et al.: Examining the threat of ChatGPT to the validity of short answer assessments in an Undergraduate Medical Program. *J. Medi. Edu. Curri. Develop.* **10** (2023). <https://doi.org/10.1177/23821205231204178>
19. Alkaissi, H., McFarlane, S.I.: Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* **15**(2), e35179 (2023). <https://doi.org/10.7759/cureus.35179>
20. Rajabi, P., Taghipour, P., Cukierman, D., Doleck, T.: Exploring ChatGPT's Impact on Post-secondary Education: A Qualitative Study. In *Proceedings of the 25th Western Canadian Conference on Computing Education (WCCCE 2023)*, May 4–5, 2023, University of British Columbia, Vancouver, Canada. *ACM International Conference Proceeding Series*, Article 9 (2023). <https://doi.org/10.1145/3593342.3593360>
21. van Lent, M., Schmorow, D.: The applications of artificial intelligence to education and training. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* **19**(2), 127–128 (2022). <https://doi.org/10.1177/15485129221088717>
22. Xia, Q., et al.: A scoping review on how generative artificial intelligence transforms assessment in higher education. *Int. J. Edu. Technol. High. Edu.* **21**(40) (2024). <https://doi.org/10.1186/s41239-024-00468-z>
23. Naidu, K., Sevnarayan, K.: ChatGPT: An ever-increasing encroachment of artificial intelligence in online assessment in distance education. *Online J. Comm. Media Technol.* **13**(3) (2023). <https://doi.org/10.30935/ojcm/13291>
24. Jeon, J., Lee, S.: Large language models in education: a focus on the complementary relationship between human teachers and ChatGPT. *Educ. Inf. Technol.* **28**(12), 15873–15892 (2023). <https://doi.org/10.1007/s10639-023-11834-1>



AI Ethics in Higher Education: A Review of Ethical Challenges

Laercio Cruvinel Júnior^{1,2} , Orrillo Héctor Ascama^{1,2,3} ,
and Mário Marques da Silva^{1,2} 

¹ Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa,
1169-023 Lisboa, Portugal

{lcruvinel, horrillo, mmsilva}@autonoma.pt

² Autónoma TechLab, 1169-023 Lisboa, Portugal

³ Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal

Abstract. This paper investigates the ethical integration of artificial intelligence (AI) in higher education, addressing the urgent need for institutions to adopt frameworks that ensure responsible use. While AI offers transformative potential on personalized learning, administrative efficiency, and enhanced teaching, it also introduces challenges related to data privacy, algorithmic bias, academic integrity, and student autonomy. Through a critical conceptual analysis supported by key examples from existing practices and scholarly literature, this paper identifies the core ethical risks associated with AI deployment in universities. It then proposes a multi-stakeholder institutional framework grounded in transparency, accountability, and equity, designed to guide ethical AI adoption across educational contexts. The paper contributes a structured model that integrates AI literacy, inclusive governance, and proactive policymaking as essential pillars for sustainable implementation. By aligning technological innovation with educational values, the work offers practical strategies for institutions seeking to use AI responsibly and ethically. This contribution is aimed at educators, administrators, and policymakers who must navigate the growing intersection of digital technologies and academic ethics.

Keywords: AI · Ethics · AI Ethics · Higher Education · AI Challenges

1 Introduction

Artificial intelligence is reshaping higher education—changing how students learn, educators teach, and how institutions function. From adaptive learning and automated grading to generative tools like ChatGPT, AI offers benefits such as personalized instruction, administrative efficiency, and research support. Yet, it also raises concerns, including data privacy, algorithmic bias, academic integrity, and reduced human agency due to opaque decision-making systems.

1.1 Definition of Ethical

Ethical. In this context, refers to the responsible and principled development, deployment, and use of artificial intelligence technologies in ways that uphold human rights, dignity, and fairness. It involves ensuring that AI systems respect privacy, promote transparency and accountability, avoid harm and discrimination, and support informed decision-making by all stakeholders—including students, educators, and administrators. An ethical approach prioritizes inclusivity, autonomy, and justice, aligning AI use with the core values and educational missions of academic institutions. This paper argues for comprehensive, inclusive institutional frameworks to guide ethical AI integration. These should ensure transparency, accountability, fairness, and user autonomy, supporting both innovation and core educational values.

1.2 Methodology

This paper employs a narrative literature review and conceptual analysis methodology. We selected academic and grey literature published from 2022 onward, focusing on the intersection of AI and ethics in higher education. Sources were retrieved through IEEE Xplore, Scopus, and Google Scholar, using keywords such as ‘AI ethics’, ‘higher education’, and ‘algorithmic bias’. The analysis was guided by thematic synthesis across policy, pedagogy, and technological implementation.

1.3 Structure

The remainder of this paper is structured as follows. Section 2 reviews the current landscape of AI in higher education, including its benefits and risks. Section 3 highlights some opportunities and advantages of AI in higher education. Section 4 identifies the ethical and practical challenges associated with the adoption of AI and explores the policy and pedagogical gaps that hinder responsible AI use. Section 5 presents the identified best practices for ethical AI integration in educational contexts. Section 6 discusses the trade-offs and tensions inherent in balancing innovation with ethical responsibility. Finally, Sect. 7 concludes with a summary of key insights and recommendations for stakeholders in education.

2 Background and Related Work

AI has become a transformative force in education, reshaping traditional paradigms and offering innovative solutions for teaching, learning, and administration [19, 22]. Tools like ChatGPT exemplify how AI can provide adaptive tutoring, streamline assessment, and enhance student engagement [5, 17]. AI’s integration in education extends beyond operational benefits; it prompts a rethinking of pedagogical practices and the role of human educators [2, 9]. The opportunities provided by AI, such as personalized learning experiences and data-driven insights, are juxtaposed with risks like diminished critical thinking, data privacy concerns, and biases embedded in AI systems [1, 16]. These

concerns are amplified in higher education, where critical reasoning and ethical decision-making are paramount [19]. This section provides an overview of current research on AI in higher education, highlighting key findings, ethical concerns, and proposed solutions from selected studies, as summarized in Table 1.

Table 1. AI in Higher Education

Study	Focus Area	Ethical Issues Identified	Proposed Solutions
Study 1 [1]	AI in Learning	Biases, ethical and privacy issues, human interactions, substitute for human judgments	Leverage individual, collaborative creative thinking and problem solving , integrate AI with an ethical and cautious approach
Study 2 [2]	AI in Management, Teaching, Learning	Ethical and humanistic aspects, security and privacy, infrastructure, over-reliance , decrease in direct communication	Ethical/humanistic aspects, necessary resources, privacy principles and ethics, training
Study 3 [16]	AI in Teaching, Learning, Evaluation	High upfront costs, data privacy concerns, technical limitations, system security, privacy, explainability, complexity	Robust AI systems, supportive policies, train, maintain human oversight, AI education platform, assessment frameworks
Study 3 [5]	AI in Instructional Design, Learning, Assessment	Cheating, data privacy, algorithmic bias, misinformation and validation, long-term learning outcomes	Cautious deployment, data protection, transparent disclosure, literacy, proactive measures, collaboration
Study 4 [21]	AI in Teaching, Learning, Research, Administration, Policies	Ethical integration, plagiarism, dependence on AI, cost, social and ethical risks, unbalanced integration	Ethical use and development, train with diverse datasets, usage guidelines, digital literacy education
Study 5 [19]	AI in Training, Research, Teaching, Learning, Pedagogy	Practical application, inequalities, risk of dangerous uses, growing ethical debate, academic integrity issues	Pedagogical intervention strategies, new teaching and assessment strategies, train faculty on AI
Study 6 [22]2024	AI in Learning, Teaching, Ethics	Privacy, overreliance on technology, potential biases , need for transparency, fairness, ethical implications of AI in data collection/decision-making, lack of research/critical thinking	Integrate core literacies , use tech responsibly/ethically , address thoughtful/responsible use, augment teachers, enhance learning, transparent AI use, responsible use, trusting environment, balance AI with human input

2.1 AI in Learning, Teaching, and Education Management

AI is widely used to enhance learning through personalized, adaptive tools like tutoring systems, chatbots, and platforms such as ChatGPT. As shown in Study 1 (Table 1), these systems support decision-making and customization. However, they also pose ethical risks—bias, privacy concerns, and potential declines in social and critical thinking skills. Researchers urge cautious, ethical integration that fosters student creativity without over-reliance on automation.

Generative AI, such as large language models like ChatGPT, broadens educational use by offering instant answers, personalized feedback, and content creation (Study 3). However, risks like cheating, misinformation, and reduced instructor roles persist. Researchers stress transparent use, strong data protection, and promoting students' critical thinking to mitigate these concerns.

AI has streamlined teaching through automated grading, personalized feedback, and interactive content. Study 2 examines smart platforms and AI-supported teaching, though concerns persist about diminishing humanistic education, data security, and weakened teacher-student interaction. Scholars advocate using AI to assist—not to replace—human judgment, along with ongoing training, ethical dialogue, and strong data policies.

In administration, AI supports student management, resource allocation, and analytics. Study 4 shows how digital frameworks and chatbots enhance efficiency but raise issues around data privacy, algorithmic transparency, and information security.

2.2 Ethical Concerns and Proposed Solutions

Although AI's ethical challenges in education are well-known, strategies that integrate policy, pedagogy, and technology use remain underdeveloped. While many sources highlight ethical issues and offer recommendations, they often treat the connection between these elements as a goal rather than an established, integrated approach.

Points that highlight the lack of integration:

1. Some authors mention the need to **incorporate cognitive capabilities and AI appropriately**, integrating AI with an **ethical and cautious approach** [1]. However, how this pedagogical integration relates to policy or ethical technological design is not always detailed as a cohesive strategy.
2. One study underlines the need for a comprehensive approach that **amalgamates technology, algorithms, legal frameworks (policy) and ethical considerations** to overcome challenges such as privacy and security [16]. This suggests that these elements may not be properly integrated at present.
3. Realizing the potential of AI requires **transformations in educational institutions and pedagogies, teacher training, and robust data governance (policy)** [9]. The need for **harmonious collaboration between humans and smart technologies** to build a humanistic and responsible educational ecosystem is recognized. The very call for this “collaboration” and “transformations” implies that full integration has not yet been achieved.
4. The **absence of a code of ethics and ethical guidelines** for the use of AI in education is explicitly indicated in a study [21]. There is also a **scarcity of literature** addressing

the ethics of AI in education and a **deficiency in “explainability,”** making it difficult to trust and understand [16].

5. The need to develop policies and guidelines **to govern the use of AI in teaching and learning is recognized** [21]. In addition, it is suggested that institutions should ensure that **digital literacy education, including AI tools, is part of the curriculum**, linking pedagogy, policy and technology, but its integrated implementation is challenging [19].
6. Several authors point to the need to **co-create a shared understanding** of how to maximize the integration of these tools into academic practice in an ethical way. The need to “co-create” implies that this understanding and integration are not yet universal or fully developed.
7. A recommendation to incorporate a **course unit on AI** into all courses would aim to integrate AI ethics into pedagogy and curriculum (policy), but its widespread implementation and impact on the use of technology still require research [11].
8. It is suggested that **educational transformations require education management to redefine policy orientations** to facilitate spaces where teachers can analyze practices and create new didactic proposals that incorporate technological developments, understanding their potential and difficulties. This is a direct call for integration between politics, pedagogy, and technology/ethics, implying that such integration is neither automatic nor easily achievable [11].
9. It is proposed that the growing ethical debate on AI “should also be **installed in educational institutions**”, thus integrating the ethical discussion into the institutional and pedagogical environment, rather than being dealt with only at the level of foreign policy or technological design [11].
10. The need for a **responsible and ethical application of AI** implies integrating technological advances with fundamental literacies [14], which highlights the crucial link between technology, ethics and pedagogy. However, addressing this integration effectively remains an area of development.

In summary, while ethical concerns shape the AI-in-education debate, sources consistently call for integrated approaches aligning policies, pedagogy, and tech design to address key challenges—lack of guidelines, limited AI literacy, and regulatory uncertainty. This paper responds by proposing a comprehensive ethical framework that balances innovation with responsibility.

3 Opportunities and Advantages of AI in Higher Education

AI can be broadly categorized into two types: **traditional AI**, which relies on rule-based algorithms and predictive models, and **generative AI**, which uses advanced machine learning techniques to create new content or solutions autonomously. This section explores how both types of AI contribute to the educational landscape.

3.1 Adaptive Learning Platforms

Adaptive learning platforms are a key AI application in education, offering personalized experiences by adjusting content based on student interactions and performance [4].

These systems identify learning gaps, suggest resources, and tailor instruction in real time. They may provide extra practice to struggling students or accelerating progress for advanced ones. This supports varied learning styles and fosters engagement. Pedagogical models like PAIR (Problem, AI, Interaction, Reflection) show how AI can enhance critical thinking and prepare students for future workforce demands [11].

3.2 Research Support

AI has also become an indispensable tool for academic research, offering significant support in various stages of the research process. Traditional AI tools, such as data mining and predictive analytics, allow researchers to process vast datasets quickly, identifying trends, correlations, and outliers that may be difficult to detect manually. These capabilities accelerate hypothesis generation and facilitate more rigorous data analysis.

Generative AI, including large language models (LLMs) like ChatGPT, has further expanded research possibilities by assisting in co-writing, drafting literature reviews, and even suggesting experimental designs. Researchers can use generative AI to produce initial drafts, refine language, or automate the synthesis of complex concepts from multiple sources. AI facilitates complex data analysis, hypothesis generation, and even co-authoring scientific work, accelerating the pace of discovery and fostering interdisciplinary collaboration [10, 11].

3.3 Administrative Efficiencies

Beyond learning and research, AI has improved educational administration by automating tasks like enrollment, scheduling, attendance, and grading, allowing staff to focus on higher-level duties. Generative AI adds value through NLP tools such as auto-replies, personalized messaging, and chatbot help desks offering 24/7 student support. These technologies enhance efficiency and student experience [2]. However, their use must align with ethical standards and institutional policies to ensure transparency, accountability, and user autonomy. The next section addresses the ethical and practical challenges of AI integration in education.

4 Ethical and Practical Challenges

Despite the transformative potential of artificial intelligence (AI) in higher education, its integration is plagued with ethical and practical challenges that require careful consideration – for example, **data privacy**, **algorithmic bias**, and **academic integrity**. The use of sensitive data raises security issues, while biases in AI models may perpetuate inequalities. Additionally, AI in grading risks undermining qualitative assessments of creativity and critical thinking without proper oversight. Responsible monitoring is essential to address these challenges [7, 10], and AI benefits must be balanced against ethical concerns, such as the potential for misinformation or over-reliance on AI-generated text.

The work from Slimi and Carballido [20] is a good start for examining the ethical challenges of Artificial Intelligence (AI) in higher education. It focuses on **biased algorithms**, **AI-driven decision-making**, and **human displacement**. Through an analysis

of seven global AI ethics policies, it emphasizes the need for **fairness**, **transparency**, and **accountability** in AI deployment to mitigate risks while maximizing its potential benefits. The study highlights the importance of collaboration among stakeholders to ensure **equitable** and **ethical use** of AI in academic settings.

The high computational demands of AI—especially large language models and deep learning systems—raise concerns about **energy use** and **environmental impact**. Training and deploying these models consume significant resources, contributing to carbon emissions and competing for fresh water. To align with sustainability goals, institutions should adopt energy-efficient AI models, promote responsible tech usage, and embed environmental considerations into digital strategies [14].

Together, these challenges highlight the importance of a balanced approach to AI integration in higher education—one that leverages AI's potential while protecting user rights, promoting fairness, and minimizing environmental impact.

5 Best Practices for Ethical AI Integration

To ensure that AI is used responsibly, institutions should adopt best practices that balance innovation with ethical responsibility. This section presents a comprehensive framework of best practices for ethical AI integration. Each category is explored with real-world examples to illustrate how these principles can be effectively applied in educational contexts.

One should note, however, that incorporating AI into the process of higher education requires time and creative attention of all people involved and that leadership is responsible for providing space and time for those learning activities to take place. Faculty, staff, and students should not be expected/required to take on this additional workload without just compensation.

1. **Ethical Implementation:** Ensure transparency, fairness, accountability, and sustainability in AI systems by mitigating algorithmic biases, prioritizing data privacy and discussing strategies to reduce environmental impact. Always disclose the AI tools used in content creation to maintain academic integrity. For instance, a university may use a clear data privacy policy for its AI-driven learning platforms, allowing students to explicitly consent to data collection with an option to opt in.
2. **AI Literacy:** Equip educators, students, staff, and administrators with a comprehensive understanding of AI's capabilities, limitations, and ethical implications through interdisciplinary courses and institutional training programs. AI literacy should include not only concepts and usage, but also critical awareness of the distinct ethical considerations associated with various types of AI systems—such as rule-based expert systems, adaptive AI, and generative AI. Developing this nuanced understanding enables all stakeholders to make informed decisions, engage responsibly with AI tools, and contribute to the ethical governance of AI in education. Today, universities and educational platforms worldwide are offering targeted courses and tutorials to support this essential knowledge base.
3. **Personalized Learning:** Use AI tools to support, not replace, traditional teaching. Adaptive platforms can enhance engagement by tailoring content to individual needs, while clear communication on AI's role builds trust.

4. **Assessment Practices:** Balance AI automation with human oversight in grading to ensure fairness. Integrate AI with interactive assessments to better evaluate critical thinking and creativity. The use of AI-powered grading assistants can provide initial feedback on student submissions, but final grades should be assigned by human instructors.
5. **Human-AI Collaboration:** Leverage AI to assist, not replace, educators. Virtual assistants can handle routine tasks, freeing faculty for more complex academic responsibilities, while supporting research through data analysis.
6. **Accessibility and Equity:** Inclusive AI tools reduce disparities and bridge the digital divide. Provide access to devices and training, ensuring underrepresented groups benefit equally.
7. **Policy Evaluation:** Regularly review AI policies to address emerging challenges. Involve diverse stakeholders to keep practices ethical and relevant. For example, conduct annual reviews of the AI ethical guidelines, incorporating feedback from students, faculty, and external experts, as an iterative process.
8. **Responsible Innovation:** Promote critical thinking and experiential learning to ensure AI supports—not replaces—human judgment in complex decisions. Encourage responsible innovation by developing AI tools that are transparent, fair, and aligned with core educational values.

These practices balance AI's transformative potential with ethical responsibility, empowering institutions to foster innovation while promoting equitable and inclusive education.

6 Discussion

The rapid integration of artificial intelligence (AI) in higher education presents both unprecedented opportunities and significant ethical challenges. As institutions strive to leverage AI for enhanced learning, teaching, and administration, they must navigate a complex landscape of competing values and priorities. This section explores the key tensions and trade-offs associated with AI integration and clarifies the roles of various stakeholders in ensuring ethical and responsible AI use in educational settings.

6.1 Tensions and Trade-Offs

AI offers substantial benefits to higher education, including personalized learning experiences, streamlined administrative processes, and enhanced research capabilities. However, these advantages are accompanied by ethical and practical trade-offs that institutions must carefully manage:

Innovation vs. Academic Rigor

AI can automate content creation, generate instant feedback, and even co-author research papers. However, this ease of content generation risks undermining academic rigor, as students may rely on AI-generated solutions rather than engaging in deep, critical thinking [15]. For instance, using AI-powered essay generators without proper guidance may lead to superficial learning and academic dishonesty. Institutions must ensure that AI

tools are used to supplement, rather than replace, traditional learning methods, preserving the value of critical analysis and original thought. As highlighted by Woerner et al. [22], unchecked reliance on generative AI tools can erode student agency, potentially leading to superficial learning outcomes. This concern aligns with our identified tension between automation and academic rigor.

Automation vs. Autonomy

AI-powered administrative tools can efficiently manage student data, automate course scheduling, and streamline student support services. However, excessive automation may erode user autonomy, making students and educators feel like passive recipients of algorithmic decisions [12]. For example, automated grading systems can efficiently evaluate large volumes of student work, but if used without human oversight, they may misinterpret creative answers or fail to recognize nuanced understanding [6]. Balancing automation with human oversight is essential to maintaining user agency.

Efficiency vs. Ethics

AI-driven systems can dramatically increase efficiency, saving time and resources for both educators and administrators. However, this efficiency must not come at the expense of ethical principles [13]. For instance, AI chatbots can provide instant student support, but if not properly monitored, they may inadvertently collect sensitive data without adequate privacy safeguards. Institutions must ensure that AI systems are designed and operated transparently, with clear data protection policies and user consent mechanisms [3].

Access vs. Equity

While AI has the potential to democratize education by providing personalized learning to a wide audience, it can also deepen existing inequalities if access is uneven [8]. Students in well-funded institutions may benefit from advanced AI tools, while those in under-resourced schools may be left behind. To prevent this digital divide, institutions must ensure that AI-enhanced learning tools are accessible to all students, regardless of their socioeconomic background [18].

6.2 The Role of Stakeholders

Ethical and responsible AI integration in higher education requires collaboration among all stakeholders—students, educators, administrators, and policymakers, each playing a distinct yet complementary role. By working together, they can develop, implement, and sustain AI practices that serve the best interests of education and society.

- **Students:** Students must be active participants in ethical AI use, developing critical thinking skills that allow them to evaluate AI-generated content and avoid over-reliance on automated tools. Institutions should provide AI literacy training, teaching students how to use these tools responsibly and understand their limitations.
- **Educators:** Educators play a critical role in modeling ethical AI use, demonstrating how to engage with AI tools for content creation, research, and assessment without compromising academic integrity. They should maintain a balance between using AI

for efficiency and fostering student creativity and independent thinking. Additionally, educators should participate in ongoing professional development to stay informed about the latest AI advancements and ethical considerations.

- **Administrators:** Administrators are responsible for setting institutional policies that govern AI use, ensuring compliance with data protection laws, and promoting transparency in AI-driven decision-making. They should conduct regular policy evaluations, maintain clear data management protocols, and foster an inclusive digital environment where all students can benefit from AI tools.
- **Policymakers:** Policymakers must collaborate with students, educators and staff to establish clear regulatory frameworks that protect user privacy, prevent algorithmic bias, and promote ethical AI development. They should work with educational institutions, technology developers, and ethicists to create standards and guidelines for AI use in education. These policies should prioritize user rights, data protection, and fairness while enabling innovation.
- **Developers:** Developers of educational AI tools have a responsibility to design systems that are transparent, secure, and user-friendly. They should prioritize fairness in algorithm design, regularly audit their models for bias, and ensure that AI systems can explain their decisions to users. Developers must also work closely with educators to ensure that their tools are pedagogically sound and aligned with ethical values.

7 Conclusions

Artificial intelligence (AI) has the potential to transform higher education by enabling personalized learning, enhancing teaching, and streamlining administration. Yet this transformation brings ethical and practical challenges, including data privacy, algorithmic bias, academic integrity, and AI's environmental impact. Institutions must address these to ensure responsible and sustainable AI integration.

Achieving this requires collective action from all stakeholders—students, educators, administrators, policymakers, and developers. Educators should model ethical AI use and promote critical thinking. Administrators must set clear policies for data use, transparency, and consent while ensuring equitable AI access. Policymakers must create regulatory frameworks that protect user rights without hindering innovation. Developers should embed fairness, accountability, and explainability into AI design, with regular audits to prevent bias.

The ethical use of AI should be supported by ongoing dialogue, adaptive policies, and a commitment to transparency. Institutions must review AI policies regularly, gather stakeholder input, and remain alert to new ethical concerns. Through a proactive, collaborative approach, higher education can embrace AI's potential while upholding core human values and fostering inclusive innovation.

To ensure that AI integration in higher education remains ethically grounded and pedagogically effective, it is essential that institutional strategies move beyond normative aspirations and be rooted in evidence-based practices. This entails systematically evaluating the impact of AI tools on student learning outcomes, equity, and well-being. By fostering a culture of critical inquiry and empirical assessment, universities can develop adaptive policies that respond to the evolving challenges of AI while upholding their educational mission and ethical responsibilities.

References

1. Al Maqbalia, S.B., Md Tahir, N., Hussain Alhazmi, O.: AI opportunities and risks for students' decision-making skill. 2024 IEEE 14th International Conference on Control System, Computing and Engineering (ICCSCE), pp. 321–325 (2024). <https://doi.org/10.1109/ICCSCE61582.2024.10696240>
2. Alomran, H.M., Alhazmi, O.H.: Artificial Intelligence and Its Effectiveness in Modern Teaching. 2023 International Conference on Data Science, Agents, and Artificial Intelligence (ICDSAAI), pp. 2–7 (2023). <https://doi.org/10.1109/ICDSAAI59313.2023.10452470>
3. Contreras, M.R., Jaimes, J.O.P.: AI Ethics in the Fields of Education and Research: A Systematic Literature Review. 2024 International Symposium on Accreditation of Engineering and Computing Education (ICACIT), pp. 1–6. Bogota, Colombia (2024). <https://doi.org/10.1109/ICACIT62963.2024.10788651>
4. Dutta, S., Ranjan, S., Mishra, S., Sharma, V., Hewage, P., Iwendi, C.: Enhancing educational adaptability: a review and analysis of AI-driven adaptive learning platforms. 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM), pp. 1–5. Noida, India (2024). <https://doi.org/10.1109/ICIPTM59628.2024.10563448>
5. Emdad, F.B., Ravuri, B., Ayinde, L., Rahman, M.I.: ChatGPT, a Friend or Foe for Education? Analyzing the User's Perspectives on The Latest AI Chatbot Via Reddit. 2024 International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), pp. 20–27 (2024). <https://doi.org/10.1109/IATMSI60352.2024.10477859>
6. Gustilo, L., Ong, E., Lapinid, M.R.: Algorithmically-driven writing and academic integrity: exploring educators' practices, perceptions, and policies in AI era. *Int. J. Educ. Integr.* **20**, 3 (2024). <https://doi.org/10.1007/s40979-024-00153-8>
7. Halvorson, O.: AI in academia: policy development, ethics, and curriculum design. *SLIS Stud. Res. J.* **14**(1), 1–7 (2024)
8. Hosltein, K., Doroudi, S.: Equity and Artificial Intelligence in Education: Will “AIED” Amplify or Alleviate Inequalities in Education? *arXiv:2104.12920v1[cs.HC]* (2021). <https://doi.org/10.48550/arXiv.2104.12920>
9. Indrayani, N., et al.: Artificial Intelligence: Investigating the Impact of Teacher's Awareness, Perception, and Learning Motivation on Learning Evaluation. 2024 International Conference of Science and Information Technology in Smart Administration (ICSINTESA), pp. 33–40 (2024). <https://doi.org/10.1109/ICSINTESA62455.2024.10747867>
10. Jacques, P.H., Moss, H.K., Garger, J.: A Synthesis of AI in Higher Education: Shaping the Future. *J. Behav. Appl. Manage.* **24**(2), 103–111 (2024). <https://doi.org/10.21818/001c.122146>
11. Maya, H.R.R., Valdes, A.B.S.: Empowering critical thinking through AI: the PAIR model's impact on higher education excellence. 2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2024 8th International Symposium On, pp. 1–6 (2024). <https://doi.org/10.1109/ISMSIT63511.2024.10757186>
12. McConvey, K., Guha, S.: This is not a data problem: Algorithms and Power in Public Higher Education in Canada. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 16, pp. 1–14 (2024). <https://doi.org/10.1145/3613904.3642451>
13. Mishara, P.: The Ethical Implications of AI in Education: Privacy, Bias, and Accountability. *J. Info. Edu. Res.* **4**(2) (2024). <https://doi.org/10.52783/jier.v4i2.1827>
14. Moreira, F., Santos-Pereira, C., Lobo, C.A.: Innovations in sustainable Higher Education: Generative AI and the future of teaching and learning. In: Moreira, F., Santos-Pereira, C.,

- Lobo, C.A. (eds.) *Innovations in Sustainable Higher Education: Generative AI and the Future of Teaching and Learning*. Journal of Infrastructure, Policy and Development, (Special Issue). Repositório Institucional UPT (2024). <https://Hdl.Handle.Net/11328/5970>
15. Octaberlina, L.R., Muslimin, A.I., Chamidah, D., Surur, M., Mustikawan, A.: Exploring the impact of AI threats on originality and critical thinking in academic writing. *Edelweiss Appl. Sci. Technol.* **8**(6), 8805–8814 (2024). <https://doi.org/10.55214/25768484.v8i6.3878>
 16. Pan, M., Wang, J., Wang, J.: Application of artificial intelligence in education: opportunities, challenges, and suggestions. 2023 International Conference on Information Technology in Medicine and Education (ITME), pp. 623–628 (2023). <https://doi.org/10.1109/ITME60234.2023.00130>
 17. Patel, R., et al.: ChatGPT in the classroom: a comprehensive review of the impact of ChatGPT on modern education. 2023 International Conference on Intelligent Systems and Embedded Design (ISED), pp. 150–157 (2023). <https://doi.org/10.1109/ISED59382.2023.10444568>
 18. Roshanaei, M., Olivares, H., Lopez, R.: Harnessing AI to Foster Equity in Education: Opportunities, Challenges, and Emerging Strategies. *J. Intel. Learn. Sys. Appl.* **15**, 123–143 (2023). <https://doi.org/10.4236/jilsa.2023.154009>
 19. Ruíz, N., Martínez, L., Laplagne, M.C.: The impact of artificial intelligence in higher education. South American Conference on Visible Light Communications (SACVLC), pp. 153–158 (2023). <https://doi.org/10.1109/SACVLC59022.2023.10347759>
 20. Slimi, Z., Carballido, B.V.: Navigating the ethical challenges of artificial intelligence in higher education: an analysis of seven global AI ethics policies. *TEM Journal* **12**(2), 590–602 (2023). <https://doi.org/10.18421/TEM122-02>
 21. Venter, I.M., Cranfield, D.J., Blignaut, R.J., Achi, S., Tick, A.: Conversational AI in higher education: opportunities, challenges, and ethical considerations. 28th IEEE International Conference on Intelligent Engineering Systems (INES), pp. 000195–000202. Gammarth, Tunisia (2024). <https://doi.org/10.1109/INES63318.2024.10629104>
 22. Woerner, J.H.R., Turtova, A.P., Lang, A.S.I.D.: Transformative potentials and ethical considerations of AI tools in higher education: case studies and reflections. SoutheastCon 2024, pp. 510–516 (2024). <https://doi.org/10.1109/SOUTHEASTCON52093.2024.10500042>



Just Hallucinations? The Problem of AI Literacy with a New Digital Divide

Eleonora Bassi¹  and Ugo Pagallo²  

¹ Nexa Center, Polytechnic of Turin, Italy

² Department of Jurisprudence, University of Turin, Turin, Italy
ugo.pagallo@unito.it

Abstract. The paper examines the normative and epistemic challenges of AI literacy with problems of applied ethics, open issues in the legal domain, and how they are related to digital divide concerns. The lack of skills, knowledge and understanding of how technology functions may depend on previous drivers of social inequality and digital division, and yet, today's AI illiteracy can trigger a new digital divide much harsher than the first wave of the late 1990s and early 2000s. Threats of AI illiteracy are systemic, exponential, cumulative, and can be worsened by the lawmaker's response, or lack thereof. The open issues of the digital divide are unpromising. Although LLMs hallucinations are only a part of a more intricate problem, it would be a first great step forward to increase the number of people capable of understanding they are just dealing with such hallucinations.

Keywords: Applied Ethics · Artificial Intelligence · Digital Divide · Hallucination · Information and Communication Technology · Law · Literacy

1 Introduction

The Cambridge dictionary chose “hallucination” as the word of the year in 2023. The traditional definition was complemented with the notion of AI systems — in particular, Large Language Models or, in the jargon of EU law, General Purpose AI (‘GPAI’) — that produce false information about the states of reality. The popularity of the neologism has gone hand-in-hand with a growing attention to the issues of AI literacy among scholars [1–3]; and institutions [4–6].

AI literacy has to do with the ability of individuals to understand, assess, and interact with AI technologies. The notion entails multiple cognitive competences and different social roles, such as those of developers and lawmakers, retailers and end-users of the technology. The challenges of AI literacy may concern as many different problems as public education and individual awareness, equity and inclusion, bias and fairness, people's privacy and data protection. AI illiteracy often leads to less well-informed users, or to human-computer interaction harms that depend on users overvaluing the capabilities of the technology, therefore using it in unsafe ways. Several examples illustrate this threat, from Chat GPT-generated students' theses to attorneys providing memorandum

of law on the basis of non-existent cases, quotes, and holdings, as discussed before the Southern District of New York in May 2023 (case *Mata v. Avianca*, no. 22-cv-01461). AI literacy thus entails a complex set of epistemic and normative challenges that regard, on the one hand, alethic information concerning the recognition and identification of the “states of reality” (strict epistemic challenges) that go together with the formation of cognitive processes and their replicability (educational challenges of AI literacy). On the other hand, the normative challenges of AI literacy concern the transformative effects of technology and its unfair outcomes that may affect human autonomy as much as the stability of the political institutions with their legitimate processes of change.

Against this framework, the claim of this paper is that a new digital divide, much harsher than the first wave of the late 1990s and early 2000s, may follow or be exacerbated by today’s AI illiteracy, in particular, of LLMs- or GPAI-related illiteracy. To substantiate this claim, the analysis of the paper is divided into four parts. Section 2 illustrates the ways in which AI illiteracy may hinge on previous drivers of social inequality and digital division, such as educational, generational, attitudinal, geographical, or socio-economic factors that add to problems of individuals with physical disabilities, lack of digital training, legal rights, or infrastructure [7, 8]. Section 3 emphasizes the uniqueness of AI — and moreover, of LLMs or GPAI technologies [9] with their risks and challenges [10, 11] — that may not only draw on the cumulative effect of multiple waves of digital divide, but also, trigger new threats of illiteracy [12, 13]. Section 4 complements the analysis of applied ethics, i.e. the moral dilemmas that arise with the social and political impact of technological innovation, with the response of lawmakers. Attention is drawn to how the law has intended to tackle the challenges of AI literacy through provisions on risk assessment, informed oversight, recruitment of talent, general awareness, education, media literacy, and critical thinking [14]. Section 5 focuses on the open issues of the field and the challenges of the future [15, 16].

The overall aim is to stress that the speed of technological innovation, in particular, with the development of new GPAI models and systems, together with some problems related to the lawmakers’ response to the challenges of AI literacy, may worsen the new digital divide which is looming on the horizon with new representational harms, information hazards, misinformation threats, or people overvaluing the capabilities of the technology. All in all, the problem does not only concern the hallucinations of AI systems, but also, the epistemic, educational, normative and political challenges posed by individuals who are not even capable of understanding they are facing just hallucinations.

2 Mind the Gap

Technological revolutions traditionally trigger new inequalities, or exacerbate old ones [17]. The “digital age” [18], the “information age” [19], or the “fourth revolution” [20] are no exceptions. On the one hand, what is new today is that for the first time ever human societies are not only bound by the use of information and communication technologies (ICTs), but depend on such digital technologies and further, on data and information as their vital resources [21]. On the other hand, the extraordinary opportunities brought forth by the “digital revolution” have their price, which includes the increasing gap between people who have access to the technology, the internet, and digital literacy

skills, and those who do not. The digital revolution came along with a “digital divide,” according to the formula of the US president Bill Clinton in his 2000 State of the Union speech.

Inequality and division between the digital haves and have-nots have raised new problems in the moral and legal domain because the pervasive use of ICTs triggers a new generation of dilemmas that can hardly be fixed in accordance with old notions and categories. Scholars have accordingly dissected the new challenges of inequality and division in the digital era in connection with manifold factors that do not consist in mere access problems [8]. These factors include education, demography, geography, socio-economics, infrastructures, training, individual attitudes and disabilities, and the law. In the domain of applied ethics, the Ethicomp community has examined many of such moral dilemmas over the past years [22], for example, focusing on education [23], marginalization and poverty [24], or access and competence [25]. Furthermore, attention has been drawn to a key driver of inequality and division, such as individual illiteracy, since digital technologies can effectively be used and understood only on the basis of appropriate skills and knowledge. As a result, several Ethicompers have focused on the abilities to use ICT tools and media literacy [26], children literacy [27], or inter-generational inequity [28], among others.

Against this framework of normative and epistemic challenges of digital literacy, we can appreciate the further disruption brought forth by Artificial Intelligence (AI) systems, in particular, new forms of AI, such as Generative AI, Foundation Models, and Large Language Models (LLMs), which broke the news by the end of 2022 and throughout 2023. For example, ChatGPT reached an amazing 100 million monthly active users within two months of its launch on 30 November 2022. The unique challenges of AI concern how this technology can be overused, or misused in society, thereby devaluing human skills, removing human responsibility, reducing human control, or eroding human self-determination [29]. In the legal domain, the High Level Expert Group on liability and new technologies formation, set up by the European Commission in 2018, similarly stressed the crucial changes brought forth by AI that depend on the complexity, opacity, openness, autonomy, predicability, data-drivenness, and vulnerability of emerging digital technologies. Such properties clearly affect the multiple cognitive competences that regard the ability of individuals (and institutions) to understand, assess, and interact with AI technologies. In the phrasing of the Report on *Liability for Artificial Intelligence*, “each of these changes may be gradual in nature, but the dimension of gradual change, the range and frequency of situations affected, and the combined effect, results in disruption” [30].

The disruption of AI systems and models should not be grasped in isolation. In the phrasing of William B. Arthur, such a disruption is rather combinatorial and recursive [31]. For example, a current smartphone equipped with AI makes use of electrochemistry in batteries for power, electromagnetic waves for communication, and quantum tunnelling in transistors for processing. Moreover, both ICTs and AI should be properly conceived as samples of the same coin because they are all types of “third-order technology”, i.e. they set the environment in which technologies function in between further technologies [32]. Therefore, we can focus on how AI systems and models, such as LLMs, may affect people’s literacy on the basis of a methodological warning.

The effects can indeed be cumulative because AI and LLMs threats risk adding to the unsolved problems of digital literacy and the digital divide under scrutiny in this section. What should we expect as a result?

3 Mind the Increasing Gap

We noted above in the introduction that “hallucination” became one of the most popular and defining LLMs threats in 2023. Alas, this is not the only kind of problem triggered by the technology. Some scholars divide the normative challenges of LLMs into six risk fields [10]. They regard, (i) discrimination, for LLMs can perpetuate social stereotypes and biases, causing representational and allocational harms; (ii) information hazards that may infringe people’s privacy because of personal information and sensitive data leaking; (iii) misinformation hazards that can affect trust, or lead to less well-informed users; (iv) malicious uses by people with criminal intent, for example, through large-scale frauds or personalized scams; (v) human-computer interaction harms that depend on users overvaluing the capabilities of LLMs and using them in unsafe ways; and, (vi) environmental harms related to the energy power which is necessary to train and operate the technology [33].

This scenario clearly affects the epistemic and normative features of both AI literacy and social inclusion. The epistemic challenges are related to the delegation of cognitive tasks and decision-making processes to AI that can end up with inconclusive evidence leading to unjustified actions, or inscrutable evidence leading to opacity, or misguided evidence leading to biases [12]. The normative challenges concern unfair outcomes that can end up either with discrimination, or with transformative effects that challenge human autonomy [13]. The emergence of several open source AI models, such as Meta’s Llama 2 and Stable Diffusion, reinforces these trends [34]. For example, according to certain scholars [35], AI-generated fake news are more credible to human raters than human-written disinformation. Moreover, texts, images, audios, and videos generated by LLMs can be impossible to distinguish from human creation [11].

However, problems of AI literacy and hallucinations do not only regard the populace. The casual use of technology also affects some fellow colleagues. The recent case of the “vegetative electron microscopy” illustrates this trend on false information about the states of reality that LLMs preserve, reinforce, and make it almost impossible to remove from knowledge repositories. The anatomy of the hallucination starts in this case with two papers that the journal *Bacteriological Reviews* published in 1950. Apparently, in the digitizing process, the term “vegetative” from one column of text was erroneously combined with “electron” from another, therefore creating the nonsensical formula [36]. In 2017 and 2019, the “vegetative electron microscopy” popped up in two papers; in 2025, according to Google Scholar, in 22: “One was the subject of a contested retraction from a Springer Nature journal, and Elsevier issued a correction for another” [36]. Since some of the articles that mention the “vegetative electron microscopy” have hundreds of citations, some wonder whether we are dealing with fake-citations as occurred in a previous case [37].

Unsurprisingly, an increasing amount of scientific papers is devoted to the assessment of digital literacy in academic achievements [38, 39], the assessment of teachers’

digital literacy [40], and the ability of young individuals to responsibly navigate digital platforms [41]. A full array of scholars and students overvaluing the capabilities of LLMs, or using such technology in treacherous ways, or with malicious purposes, illustrate why this kind of research is urgent. Further work covers digital literacy status [42], knowledge [43], competence [44], and strategies [45]. Lack of digital literacy is indeed critical not only for educational purposes and the future of academic research in a LLMs world, but also, considering that problems of algorithmic awareness, or misinformation hazards which affect trust, are systemic and span across society, from developers to lawmakers, consumers and citizens, retailers and end-users. The illiteracy threat affects the very stability of the institutions. This political impact is crystal clear and yet, has often been overlooked in the debate.

As a result, the attention should be drawn to how technological illiteracy often affects courts, public servants and administrations, down to legislators and policy makers. Who educates the educator?

4 When Illiteracy Meets the Law

Legal regulators are often invoked to take charge of both the literacy and digital divide problems that are related to matters of access and infrastructure, usage and individual skills, or hazardous outcomes triggering inequality and division. In several cases of digital literacy, the tools of hard law — namely, the rules of the law that intend to govern social and individual behaviour hinging on the threat of sanctions — arguably fall short in coping with the multiple cognitive competences that shall provide for the ability of individuals to comprehend, evaluate, and interact with digital technologies. This is rather an aim of soft law, namely, guidelines, opinions, and recommendations that, although not legally binding, provide authoritative instructions as to how to interpret and implement the law. Legislators can hardly defeat any AI illiteracy by decree [46, 47].

This mix of soft law and hard law, on the other hand, does not only have to address the needs of private parties, but also, of public actors and institutions [48]. Lack of literacy is critical because it threatens the development of all those sectors — from the digitization of public services to the competitiveness of private companies — which require new skills, abilities, and expertise to tackle the pressing problems of today and the uncertain challenges of the future. The intervention of the legislator has thus to be reflective because the public administration represents both the subject and object of its own literacy policies. Consequently, scholars have examined the impact of higher education administrative staff's digital literacy on their intention to use digital technologies while performing their tasks [49]; the role of digital competences for job crafting in public administration [50]; the necessity of media literacy in police forces [51]; the digital literacy of courts and the judiciary [52]; down to the digital literacy of lawmakers [53].

Digital literacy can thus be understood either as an aim to be attained or as a means of regulation. Regarding the aims to be attained, focus of lawmakers is on requirements of inclusion and equality, such as individual autonomy, awareness, or proper understanding of the use of technology. Regarding the means of regulation, the attention of lawmakers concerns the challenges of literacy as a condition for the modernization and functioning of all those fields — from digital public services to the private sector — which critically

hinge on new talents, capabilities, and knowledge [14]. To grasp how all this works in practice, it is worth mentioning the ways in which legislators have intended to tackle the challenges of digital literacy brought forth by advancements of AI, especially considering that AI illiteracy is cumulative, that is, it may depend on previous drivers of social inequality and digital division with general problems of access, usage, and undesirable outcomes.

In EU law, for example, the AI Act, i.e. Reg. (EU) 2024/1689 from July 2024, intends to tackle issues of AI literacy according to four sets of legal provisions [5]. They regard (i) legal duties of providers and deployers of the technology; (ii) compliance and enforcement of the Regulation; (iii) general awareness; and, (iv) education. According to Art. 3(56) on definitions of the AI Act, literacy is understood as “skills, knowledge and understanding that allow providers, deployers and affected persons... to make an informed deployment of AI systems, as well as to gain awareness about the opportunities and risks of AI.” All previous differentiations of the law are at work with the EU regulation, such as the mix of hard law and soft law, the public administration as the subject and object of the legal provisions, and AI literacy considered as a means and an end of the legislation. Specific provisions are devoted to the deployment of AI systems in education. In the wording of Recital 56, it is “important to promote high-quality digital education and training and to allow all learners and teachers to acquire and share the necessary digital skills and competences, including media literacy, and critical thinking, to take an active part in the economy, society, and in democratic processes.”

In the US, the Biden’s administration released the *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* in October 2023 [6]. Such provisions were overruled by President Trump in the first week of his administration in January 2025 [54], and substituted with a new Executive Order on *AI Education for American Youth* in April of that year [55]. The provisions of Biden’s Executive Order on AI literacy were related to (i) human capabilities to assess risks; (ii) the aim to attain informed oversight; and, (iii) recruitment of talent in the public sector. The provisions of Trump’s Executive Order on AI education aim to “fostering interest and expertise in artificial intelligence (AI) technology from an early age to maintain America’s global dominance in this technological revolution for future generations” [55]. To attain this end, the plan is to (i) establish the White House Task Force on AI Education; (ii) implement “a Presidential AI Challenge, which will encourage and highlight student and educator achievements in AI”; (iii) create public-private partnerships to provide resources for “K-12 AI education”; (iv) prioritize the use of AI in “discretionary grant programs for teacher training”; and, (v) “protecting and preparing the American workforce for the challenges of the future.” As a response, 48 House Democrats sent a letter to the Office of Management and Budget demanding “immediate termination” of AI systems used by the Department of Government Efficiency (DOGE), since they were “specifically concerned about reports of Elon Musk and DOGE’s monitoring and sharing of federal employee and non-public federal data using AI tools” [56].

5 Open Problems

Whether and to what extent efforts of national, regional, and international legislators and policy makers will successfully address the challenges of AI literacy remains an open issue, especially considering that, for example, the AI Act of EU law will be at full speed only in August 2026. Despite such legal uncertainties, it seems fair to admit, however, that the stakes are high because of a potential vicious circle between the challenges of AI literacy and a persisting digital divide. To substantiate the assumption, attention must be drawn to four critical issues. The first problem regards how current initiatives of lawmakers on AI literacy cover only a part of the plethora of issues mentioned in the previous parts of this paper with work in applied ethics. Even admitting the success of the initiatives and regulations of the AI Act, or of Trump's Executive Orders for that matter, further problems related to new representational harms, information hazards, misinformation threats, or social distrust, should still be addressed.

The second reason for concern refers to the speed of technological innovation, in particular, of LLMs and what the EU lawmakers dub 'general purpose AI'. It speaks volumes that the European Commission simply ignored the existence of such AI models in the first draft of the AI Act in April 2021 [9]. Waiting to understand how the rules of soft and hard law will be interpreted and applied in the coming months and years, new technological advancements can be reasonably expected in this span of time, triggering a cat-and-mouse game in which lawmakers may arrive 'too late' with their regulations.

A third critical issue is related to the cumulative dynamics of technological innovation [30, 31]. Threats of new AI and LLMs illiteracy must be understood in connection with persisting problems of digital illiteracy as a powerful driver of inequality and division. Such threats may increase and reinforce each other. We are far from having solved many problems of the digital divide inherited from the first waves of digitization in the late 1990s and early 2000s [7, 8], and yet a much harsher digital divide is looming on the horizon. The more LLMs or GPAI systems trigger issues of access, usage, and outcomes — as stressed above in the third section of this paper [10, 11, 33–36] — the less we are close to solving persisting problems of digital literacy and exclusion.

Last, but not least, attention must be drawn to one of the most relevant challenges of AI literacy, that is, 'meta-cognition' [15, 16]. In addition to the pressing problems of the present, focus should indeed be on the challenges of the future, i.e. how to tackle the challenges of an uncertain future through the anticipation capabilities of AI literacy. Although some scholars claim that AI systems can be understood as both the reason and solution of this new digital divide [57], the assumption cannot be reasonably extended to the regulatory problems under scrutiny in this paper. After all, we noted that it is up to lawmakers the responsibility of how to cope with the mounting problems of AI literacy, and moreover, how to prepare citizens for new incessant technological advances.

As a result of these four sets of open issues, it follows that no one-size-fits-all approach to the challenges of AI literacy seems appropriate. The challenges of AI literacy entail multiple cognitive competences as well as different social roles of individuals and public institutions that can trigger as many diverse problems as equity and inclusion, bias and discrimination, education and awareness. Therefore, it comes as no surprise that the legislative response to the epistemic and normative challenges of AI literacy has provided for a mix of soft law and hard law, public law and private law, to tackle

those challenges either as a means or as a goal of legislation. It shall be expected that the assessment of such normative efforts should similarly be calibrated in accordance with that which should or should not be done against current legislation, or in spite of it, or whether such evaluation should limit itself to ameliorate today's state of the legal art. The law can be the solution, but also, a part of the problem. The less legal systems will be effective in addressing the new generation of literacy problems triggered by AI, LLMs, or GPAI, the more such AI systems and models will trigger issues of access, usage, and outcomes sparking a new digital divide; and yet, the more a new digital divide is fated to increase, the less probable that people understand when they deal with a digital hallucination.

6 Conclusions

The paper examined the normative and epistemic challenges of AI literacy with issues of applied ethics, open problems in the legal domain, and how they are related to digital divide concerns. Although the concepts must be distinguished — since there can be problems of literacy regardless of digital divides and, vice versa, problems of digital divide may not directly concern matters of literacy, as occurs for example with issues of gender discrimination, or lack of infrastructure — it is fair to concede that these notions often overlap and, moreover, their corresponding issues may reinforce each other. The digital gap between haves and have-nots can accrue problems of literacy, and the other way around, problems of technological literacy often go hand-in-hand with the generational, attitudinal, geographical, or socio-economic factors of the digital divide.

Against this framework, the paper scrutinized the many facets of AI illiteracy, in particular, the strict epistemic, educational, normative, and political threats brought forth by the use of LLMs and GPAI. The lack of skills, knowledge and understanding of public actors, private companies, and individuals to harness the opportunities of technology while preventing its risks is critical. The paper insisted on the properties and features of AI that make it unique for better and worse. Misuses and overuses of AI can devalue human skills, remove human responsibility, reduce human control, and erode human self-determination. Further, misuses and overuses of LLMs and GPAI can perpetuate social stereotypes and biases (discrimination), infringe people's self-determination (information hazards), affect trust and lead to less well-informed users (misinformation), facilitate criminal intents (malicious uses), or induce users to either overvaluing what technology does, or using it in unsafe ways (human-computer interaction harms).

The lack of skills, knowledge and understanding of how technology functions and what it can or cannot do may depend on previous drivers of social inequality and digital division due to the combinatorial and recursive nature of technology with the cumulative effect of multiple waves of digital divide. However, today's AI illiteracy, in particular, LLMs- or GPAI-related illiteracy can also trigger a new digital divide, even harsher than the first wave of the late 1990s and early 2000s. Considering the speed of innovation and the general purposes of the technology with its new and unique outputs, from texts to videos, images, audios, or 3D models, it follows that the gap between the haves and have-nots can dramatically increase in accordance with those factors of the digital divide, such as education and awareness, demography and geography.

The open problems of the digital divide inherited from the first two decades of the 2000s are unpromising. The paper dwelt on the differentiated response of the law between soft and hard law, private and public law, means and goals of legislation with four sets of open issues. Threats of AI illiteracy can indeed be systemic, exponential, cumulative, and even worsened by the lawmaker's response, or lack thereof. The bottom line is that LLMs hallucinations, as a false information about the states of reality, are only a part of a more intricate problem that is made of ethics and geopolitics, equality and discrimination, misinformation and division, malicious uses and information harms. All in all, it would be a first great step forward to increase the number of people capable of understanding they are just dealing with mere hallucinations.

Acknowledgments. This study was funded by the Italian national research program PRIN 2022 'Equitable Algorithms, Promoting Fairness and Countering Algorithmic Discrimination Through Norms and Technologies' (EQUAI) cod. MUR 2022KFLF3E - CUP D53D23007280006.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aoun, J.E.: *ROBOT-PROOF: Higher education in the age of artificial intelligence*. MIT Press, Cambridge, MA (2017)
2. Long, D., Magerko, B.: What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–16 (2020)
3. Yi, Y.: Establishing the concept of AI literacy. *Jahr-European Journal of Bioethics* **12**(2), 353–368 (2021)
4. UNESCO: *Futures literacy. An essential competency for the 21st century*. January 30 (2021). available at <https://en.unesco.org/futuresliteracy/about>
5. The White House, Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. October 30, Washington D.C., USA (2023)
6. AI Act, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024. Document 32024R1689, Brussels (2024)
7. Cullen, R.: Addressing the digital divide. *Online Inf. Rev.* **25**(5), 311–320 (2001)
8. Van Dijk, J.J.: *The digital divide*. Polity Press, Cambridge (2020)
9. Pagallo, U.: LLMs meet the AI Act: Who's the sorcerer's apprentice? *Cambridge Forum on AI: Law and Governance* **1**, e1 (2025)
10. Weidinger, L., et al.: Ethical and social risks of harm from language models. *2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229 (2021)
11. Nightingale, S.J., Farid, H.: AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proc. Natl. Acad. Sci.* **119**(8), e2120481119 (2022)
12. Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: Mapping the debate. *Big Data & Society* **3**(2) (2016)
13. Pagallo, U.: Algo-rhythms and the beat of the legal drum. *Philosophy and Technology* **31**, 507–524 (2018)
14. Bassi, E.: *Dati, sovranità, nuovi modelli di governance*. In: Durante, M., Pagallo, U. (eds.) *La politica dei dati. Il governo delle nuove tecnologie tra diritto, economia e società*. Milan, Mimesis (2022)

15. Yussen, S.R.: The role of metacognition in contemporary theories of cognitive development. In: Forrest-Pressley, D., Waller, G. (eds.) *Contemporary Research in Cognition and Metacognition*. Academic Press (1985)
16. Miller, R.: Futures literacy: transforming the future. In: *Transforming the future. Anticipation in the 21st century*, pp. 1–12. Routledge Taylor & Francis Group, London/New York (2018)
17. Xiao, A., Xu, Z., Skare, M., et al.: Bridging the digital divide: the impact of technological innovation on income inequality and human interactions. *Humanit Soc Sci Commun* **11**, 809 (2024)
18. Negroponte, N.: *Being digital*. Vintage, New York (2015)
19. Castells, M.: *The information age: Economy, society and culture*, 3 vol. Blackwell, Oxford (1998)
20. Floridi, L.: *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press, Oxford (2014)
21. Pagallo, U.: Good onlife governance: On law, spontaneous orders, and design. In: Floridi, L. (ed.) *The onlife manifesto: Being human in a hyperconnected era*, pp. 161–177. Springer, Cham (2015)
22. Stahl, B.C., Ess, C.M.: 20 years of ETHICOMP: time to celebrate? *J. Inf. Commun. Ethics Soc.* **13**(3/4), 166–175 (2015)
23. Fleischman, W.M.: Getting to the other side—beyond the digital divide. *Proceedings of ETHICOMP 2008*, pp. 245–255. University of Pavia, Mantua, Italy (2008)
24. Kumar, S., Dhariwal, M., Pujari, A. K., Sharan, R.: Information and communication technologies for consolidating democracy: A case study from India. *Proceedings of ETHICOMP 2008*, pp. 496–502. University of Pavia, Mantua, Italy (2008)
25. Firth, R.M., Flick, C.: Digital capital and sociotechnical imaginaries: Envisaging future home tech with low-income communities. *ETHICOMP* **2020**, 309 (2020)
26. Arias-Oliva, M., Yañez-Luna, J.: E-learning evolution: New concepts on ICT & learning. *ETHICOMP* **2013**, 25–34 (2013)
27. Asai, R., Yanagihara, S.: The hidden side of digital technologies for family use: Privacy and other related ethical issues in digital distance education and telecommuting at home. *ETHICOMP* **2022**, 447 (2022)
28. Alvarez, I., Silva, N.: Prospects of equity in post-pandemic technology education. *ETHICOMP* **2022**, 16 (2022)
29. Floridi, L., et al.: AI4People - An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Mind. Mach.* **28**(4), 689–707 (2018)
30. European Commission: Directorate-General for Justice and Consumers. Liability for artificial intelligence and other emerging digital technologies, Publications Office (2019). <https://doi.org/10.2838/573689>
31. Arthur, W.B.: *The nature of technology*. Free Press, New York (2009)
32. Floridi, L.: Technology's in-betweenness. *Philosophy & Technology* **26**, 111–115 (2013)
33. Pagallo, U., Ciani Sciolla, J., Durante, M.: The environmental challenges of AI in EU law: Lessons learned from the artificial intelligence act (AIA) with its drawbacks. *Transforming Government: People, Process and Policy* **16**(3), 359–376 (2022)
34. Dickson, B.: How open-source LLMs are challenging OpenAI, Google, and Microsoft (2023). <https://bdtechtalks.com/2023/05/08/open-source-llms-moats/>
35. Zellers, R., et al.: Defending against neural fake news (2019). <http://arxiv.org/abs/1905.12616>
36. Snoswell, A.J., Witzemberger, K., El Masri, R.: A weird phrase is plaguing scientific papers – and we traced it back to a glitch in AI training data. *The Conversation* (2025). <https://theconversation.com/a-weird-phrase-is-plaguing-scientific-papers-and-we-traced-it-back-to-a-glitch-in-ai-training-data-254463>
37. Ibrahim, H., Fengyuan, L., Zaki, Y., Rahwan, T.: Citation manipulation through citation mills and pre-print servers. *Sci. Rep.* **15**(1), 5480 (2025)

38. Holm, P.: Impact of digital literacy on academic achievement: Evidence from an online anatomy and physiology course. *E-Learning and Digital Media* **22**(2), 139–155 (2025)
39. Naddaf, M.: AI linked to explosion in low-quality biomedical papers. *Nature* **641**, 1080–1
40. Nguyen, L.A.T., Habók, A.: Tools for assessing teacher digital literacy: a review. *J. Comp. Edu.* **11**(1), 305–346 (2024)
41. Buchan, M.C., Bhawra, J., Katapally, T.R.: Navigating the digital world: development of an evidence-based digital literacy program and assessment tool for youth. *Smart Learn. Environ.* **11**, 8 (2024)
42. Peng, D., Yu, Z.: A literature review of digital literacy over two decades. *Edu. Res. Int.* **1**, 2533413 (2022)
43. Silamut, A.A., Petsangsri, S.: Self-directed learning with knowledge management model to enhance digital literacy abilities. *Educ. Inf. Technol.* **25**(6), 4797–4815 (2020)
44. Falloon, G.: From digital literacy to digital competence: the teacher digital competency (TDC) framework. *Education Tech. Research Dev.* **68**(5), 2449–2472 (2020)
45. Anthonysamy, L., Koo, A.C., Hew, S.H.: Self-regulated learning strategies in higher education: Fostering digital literacy for sustainable lifelong learning. *Educ. Inf. Technol.* **25**(4), 2393–2414 (2020)
46. Pagallo, U.: The politics of data in EU law: Will it succeed? *DISO* **1**, 20 (2022)
47. EU Parliament: Media literacy: Fostering a key civic skill in a digital information environment, 27 May. Briefing, Brussels. [https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI\(2025\)772886](https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)772886)
48. Pagallo, U., Casanovas, P., Madelin, R.: The middle-out approach: Assessing models of legal governance in data protection, artificial intelligence, and the web of data. *The Theory and Practice of Legislation* **7**, 1–25 (2019)
49. Kabakus, A.K., Bahcekapili, E., Ayaz, A.: The effect of digital literacy on technology acceptance: An evaluation on administrative staff in higher education. *J. Info. Sci.* 01655515231160028 (2023)
50. Buonocore, F., Agrifoglio, R., de Gennaro, D.: The role of digital competencies and creativity for job crafting in public administration. In: *Digital transformation and human behavior: Innovation for people and organisations*, pp. 87–97. Springer International Publishing, Cham (2020)
51. Jarolimek, S.: Using smartphones in police operations: the necessity of media literacy in police forces in the age of transparency. *European J. Sec. Res.* **5**, 105–114 (2020)
52. Djamaludin, D., Aziz, M.F., Ar-Rasyid, Y., As-Sayyis, I.A.: Assessing the impact of electronic court systems on the efficiency of judicial processes in the era of digital transformation. *Volksgeist: Jurnal Ilmu Hukum Dan Konstitusi*, 1–18 (2023)
53. Galloway, K.: A rationale and framework for digital literacies in legal education. *Leg. Educ. Rev.* **27**, 117–142 (2017)
54. The White House: Executive order on removing barriers to American leadership in artificial intelligence. January 23. Washington D.C., USA (2025)
55. The White House: Executive order on advancing artificial intelligence education for American youth. April 23. Washington D.C., USA (2025)
56. Congress of the US: House Democrats Seek Immediate Termination of DOGE’s Unauthorized Use of AI Systems, Call Out Security Risks and Potential Criminal Liability, April 16, <https://beyer.house.gov/news/documentsingle.aspx?DocumentID=6476>. Washington D.C., USA (2025)
57. Božić, V.: Artificial intelligence as the reason and the solution of digital divide. *Language Education and Technology* **3**(2), 90–103 (2023)



Ethical Aspects of Distributed Extended Reality Training

Olli I. Heimo^(✉)  and Teijo Lehtonen 

University of Turku, Turku, Finland
{olli.heimo, teijo.lehtonen}@utu.fi

Abstract. This study explores the ethical implications of implementing Distributed Extended Reality (XR) in the context of remote education and training. As XR technologies evolve and become increasingly integrated into learning environments, particularly in technical fields requiring hands-on experience, new ethical questions arise. The research applies a deductive method within the framework of Virtuous Just Consequentialism, an extension of Moor's Just Consequentialism, to evaluate whether distributed XR systems can be implemented without amplifying existing ethical concerns. The analysis focuses specifically on the unique ethical characteristics of distributed XR.

Key findings indicate that, under certain conditions, distributed XR solutions can be deployed in ethically sound ways that do not exacerbate existing problems. In some cases, distributed XR may even offer improvements over traditional remote education by enhancing engagement, reducing resource waste, and enabling broader access to complex technical training. However, challenges related to usability, equality, and technological limitations remain, particularly for novice users and vulnerable groups. The use of a multi-ethical analytical framework proved helpful for exploring these layered issues, though its complexity can pose interpretive difficulties.

The study concludes that distributed XR holds promise as an ethical training technology, provided that implementation is carefully designed with attention to inclusive access, technological fairness, and the cultivation of virtue in both learners and systems.

Keywords: Distributed XR · Extended Reality · Ethics · Ethics of Technology · Pedagogy

1 Introduction

Extended Reality (XR) refers to a spectrum of technologies that merge physical and digital environments in ways that reshape human perception, presence, and interaction. XR encompasses Virtual Reality (VR), which immerses users in entirely computer-generated settings; Augmented Reality (AR), which layers digital content onto the real world; and Mixed Reality (MR), where physical and digital elements co-exist and interact in real time. In recent years, the term “Metaverse” has also emerged as a related—though conceptually contested—notion, often referring to persistent, networked, and immersive

virtual spaces. These technological domains are not only shaping new modalities of experience but are also prompting interdisciplinary discussions around their ethical, societal, and philosophical implications. [1–5].

The development and operation of XR systems hinge on the convergence of advanced hardware (e.g., sensors, head-mounted displays), software (e.g., rendering engines, AI components), and networked infrastructure (e.g., 5G, edge computing). These systems are inherently socio-technical in nature, involving both technological capabilities and human-centred considerations such as usability, safety, accessibility, and psychological impact. The ethical questions posed by XR are multifaceted, touching upon issues of data privacy, embodiment, manipulation of perception, and the boundaries of reality and identity in mediated environments.

A particularly significant development within the XR field is distributed XR, which enables multiple users in geographically separate locations to engage in shared immersive experiences through networked XR devices. This technological affordance opens possibilities for remote collaboration, simulation-based training, and social interaction in virtual environments that transcend physical distance. As such, distributed XR is not merely an extension of XR in technical terms, but a transformative innovation with broader societal implications. It challenges conventional assumptions about presence, co-location, and communication, and invites further inquiry into the ethical frameworks that should guide its design and deployment.

As digitalisation and the green transition reshape working life, companies face significant challenges in workforce training. Beyond investments in equipment and software, continuous skill development is essential. For manual workers, hands-on training has traditionally required physical presence, as practical skills are difficult to develop remotely. Distributed XR offers an alternative by enabling location-independent, immersive training environments with scalable access. The increasing availability of high-quality, portable XR devices lowers the technological and financial barriers to adoption. By integrating distributed XR into training programs, companies can reduce costs and environmental impact, for example, by minimising long-distance travel. However, the effectiveness of these solutions depends on their ability to align with human factors, such as pedagogical principles and the real-world demands of work.

While the potential benefits of implementing distributed XR for training and upskilling are well recognised, successful implementation requires more than technological readiness. The situated nature of learning in manual work contexts means that realism, embodiment, and interaction fidelity are critical. If virtual training environments fail to support meaningful engagement with tasks, tools, and bodily movement, the learning outcomes may be compromised. Furthermore, trainees' motivation, cognitive load, and sense of presence all influence the effectiveness of XR-based learning. These human factors must be considered from the outset, not only as add-ons to otherwise technical systems.

This study presents a deductive analysis grounded in an extended ethical framework based on James Moor's [6] theory of Just Consequentialism. The analysis draws specifically on the model of Virtuous Just Consequentialism developed by Heimo and Kimppa [7]. The research begins with a closer examination of distributed XR technology, followed by an exploration of the theoretical framework. In the analysis section, the

presented material is examined through the adopted ethical framework, and the findings indicate that, under specific conditions, distributed XR systems may be implemented without intensifying existing ethical challenges. The conclusions reflect on these results and outline directions for further research.

2 Distributed Extended Reality Training

Traditional classroom training, led by instructors and relying on physical equipment, remains a widely used and institutionally embedded approach, particularly in fields that require manual skills or the use of specialized tools. This model offers the benefits of direct supervision, real-time feedback, and the tactile experience of working with real materials. However, it is also resource-intensive, space-dependent, and often inflexible in terms of scheduling and individual pacing.

The COVID-19 pandemic accelerated the adoption of remote training methods and catalysed broader investments in online learning infrastructure. Virtual collaboration platforms such as Microsoft Teams and Zoom enabled remote interaction and communication, but their capabilities are largely limited to verbal and visual exchange. They are not designed to support the development of hands-on skills that rely on spatial awareness, motor coordination, and physical engagement.

In response, some organisations have begun to integrate XR technologies into their training programs, offering immersive and multisensory environments for learning [1, 2]. These XR-based solutions can simulate real-world tasks with high fidelity, providing users with opportunities to practice procedures, respond to simulated scenarios, and receive feedback in real time. However, these implementations have typically been local, requiring on-site infrastructure, and have often been designed as bespoke, one-off solutions. As a result, their scalability and long-term cost-effectiveness remain limited.

The potential of Distributed XR, illustrated in Fig. 1, lies in its ability to combine the immersion and realism of local XR systems with the reach and flexibility of online platforms. By leveraging portable headsets, cloud-based rendering and distribution, platform-driven content creation tools, and AI-assisted personalisation, distributed XR systems can enable synchronous or asynchronous training across locations. This makes it possible to conduct training sessions without requiring all participants to be in the same physical space, while still preserving key aspects of embodied learning and social presence.

Furthermore, distributed XR introduces new opportunities for real-time collaboration in simulated work environments. Trainees can interact with each other and with virtual objects across distances, allowing for team-based training, peer learning, and remote mentoring. The use of modular platforms also supports the creation and reuse of training content, improving scalability and reducing development overhead. These technological affordances position distributed XR as a promising approach for addressing both current training needs and future challenges in dynamic work environments.

Recent advancements in XR technology, particularly in VR headsets, have accelerated the industry's transition from experimental applications to enterprise adoption [8]. Companies such as NVIDIA and Microsoft have introduced enterprise-focused platforms that enable businesses to establish a persistent or event-based presence in virtual

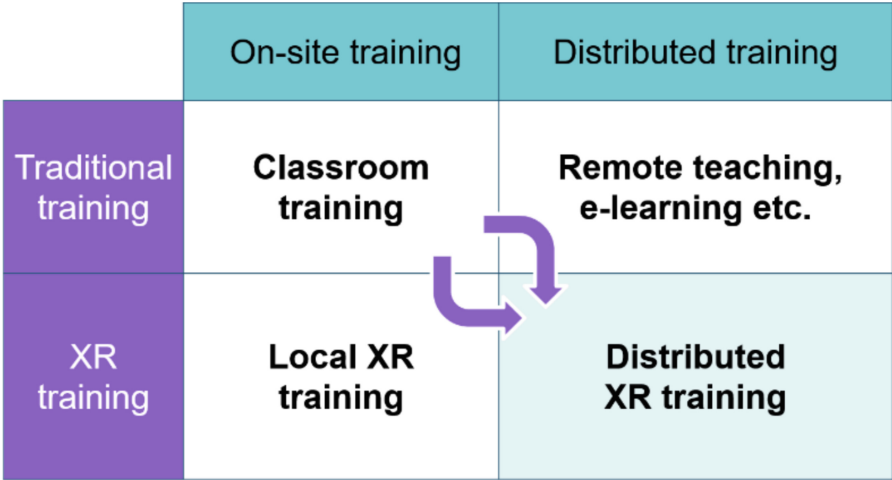


Fig. 1. Distributed XR training

environments. These platforms support new forms of communication, remote collaboration, and operational efficiency, offering organizations alternatives to traditional physical infrastructures and workflows [5]. The convergence of platform development with growing enterprise interest has created a feedback loop: as organisations identify new XR use cases, technology providers increasingly tailor their solutions to meet sector-specific needs.

Alongside platform innovation, significant hardware improvements have enhanced the usability and appeal of XR devices. The latest generation of headsets has addressed earlier barriers related to weight, resolution, battery life, and interaction modalities. Market leaders like Meta, with the Quest 3 [9], and Apple, with the Vision Pro [10], now offer devices with full-colour passthrough capabilities, allowing users to perceive and interact with their physical surroundings while engaging with digital content [11]xx. These high-definition, wide-field-of-view overlays facilitate context-sensitive interaction, blurring the boundaries between simulation and real-time operational support.

This combination of hands-free operation, spatial awareness, and real-time digital augmentation has expanded the practical use of XR beyond isolated simulation-based training. It enables integration into active work environments, where users can receive contextual guidance, manipulate digital twins, or collaborate with remote experts while remaining aware of their physical surroundings. As such, XR is no longer limited to preparatory instruction but increasingly supports just-in-time learning and decision-making in situ. These developments further underscore the relevance of distributed XR, as the same capabilities that enhance local experiences now enable meaningful participation across geographically distributed teams. [12].

Distributed XR training leverages networked XR systems to enable learning experiences that are not confined to a single physical location. Through the use of interconnected XR devices, individuals and groups can engage in immersive and interactive

training sessions while situated in different geographical places. This enables both synchronous and asynchronous forms of engagement, depending on the training scenario and technological infrastructure.

The term *training* is used here in a broad sense, encompassing a variety of learning activities: independent study, practical exercises, formative assessments, collaborative group tasks, and instructor-led sessions. These may involve direct, continuous guidance from a trainer, or on-demand assistance facilitated through embedded instructions, AI-driven agents, or contextual prompts. The distributed nature of such systems supports both formal and informal learning, allowing organizations to design training processes that align with their specific operational needs and resource constraints.

By integrating real-time communication, shared virtual environments, localisation capabilities, and AI-assisted guidance, distributed XR training offers a scalable and flexible framework for skill development across multiple sectors. Industries ranging from construction and logistics to healthcare and manufacturing can benefit from the ability to train workers in situ or remotely, without compromising on engagement or effectiveness. These systems also facilitate team-based learning, enabling geographically dispersed participants to interact in simulated work environments that closely mirror real-world conditions.

To emphasise: recent advances in network infrastructure, including the increasing availability of high-bandwidth mobile and satellite connections, now make it feasible to deploy distributed XR training even in remote or infrastructure-poor regions. Combined with the declining cost and increasing portability of XR hardware, this opens new possibilities for inclusive access to training, especially in contexts where traditional methods are logistically or economically unviable. In doing so, distributed XR may not only complement existing training models, but also transform who has access to them, where, and under what conditions.

As XR technology continues to evolve, the integration of artificial intelligence introduces new dimensions that significantly extend the potential—and complexity—of these systems. AI enhances XR by enabling adaptive and personalised user experiences, facilitating real-time decision-making, and supporting intelligent virtual agents capable of responding dynamically to human actions. These developments contribute to making XR environments more intuitive, accessible, and responsive to diverse user needs, while also improving operational efficiency and reducing dependency on human-led instruction.

AI-driven automation can support the rapid creation and customisation of XR content, optimise training flows, and allocate resources more effectively. This makes XR-based training and collaboration increasingly scalable, even in contexts where human trainers or instructional designers are not readily available. In parallel, machine learning models can analyse performance data to offer individualised feedback, adapt the difficulty of scenarios, and anticipate user behaviour—all of which can enhance learning outcomes when applied responsibly.

3 Ethical Framework

Ethical considerations surrounding XR technologies have been a subject of study for over a decade, but the rapid evolution of this field requires ongoing reassessment to address new and emerging challenges. Heimo et al. [11] contributed by categorising XR-related ethical concerns into two main categories: **direct** and **collateral** consequences. Direct issues are those that arise predictably from the adoption and integration of XR technologies, and they include well-established concerns such as privacy, surveillance, and data ownership. These issues are immediately apparent as they directly relate to the handling of user data and the potential for invasion of privacy, making them central to ethical analysis in XR. On the other hand, **collateral** issues arise more indirectly, often as a result of the limitations or weaknesses inherent in existing technological systems, or due to the inability of developers to foresee all potential consequences of their innovations. One such collateral issue is the tendency of XR technologies to be developed with a narrow user perspective in mind—typically from the standpoint of the developers themselves, which may lead to biases in design that exclude diverse user groups. This phenomenon, often described as the “pale male problem,” highlights the importance of considering inclusivity and diversity in the design and deployment of XR systems. [13].

While examining these consequences is important for understanding the ethical landscape of XR technologies, focusing solely on the outcomes provides a limited perspective. The consequences of XR adoption, while critical, must be considered within the broader context of the processes that shape these outcomes. Specifically, a more comprehensive ethical analysis must consider the values, intentions, and broader societal implications that influence the development, implementation, and user experiences of XR systems.

In this regard, the application of **virtue ethics** in the analysis of XR is both relevant and well-grounded. Virtue ethics places a strong emphasis on character formation and habitual action, prioritising the development of virtues over adherence to rigid, rule-based ethical systems. This makes virtue ethics particularly suitable for evaluating technologies like XR, which shape user behaviour and engagement over time. Unlike deontological or consequentialist approaches, virtue ethics does not focus on prescriptive rules or outcomes. Instead, it offers guidelines that encourage individuals and organisations to cultivate virtues—such as honesty, empathy, and fairness—and avoid vices—such as dishonesty or exploitation. Its normativity operates at the level of ideals, fostering the development of morally commendable character traits rather than prescribing specific behaviours. In the Aristotelian tradition, true value arises only through the consistent practice of virtues. A person does not become virtuous merely by adhering to virtuous actions reluctantly or out of duty. Instead, true virtue is the result of a long-term commitment to acting virtuously in a way that becomes natural and ingrained in one’s character. [[14], I, 9–10; 1098a15–21; 1098b5–30; 1100a31–1101a21, II, 1; 1103a31–1103b25; 1104a10–1105a16, [15]].

As Vallor [16] further argues, the moral development of individuals cannot be assessed merely by their thoughts, emotions, or beliefs. Rather, it is essential to examine the actions they consistently engage in and determine whether those actions cultivate virtues or reinforce vices. In the context of XR, this implies that ethical evaluation should

not solely focus on the immediate outcomes of XR applications, but also on how these technologies shape the long-term behaviours and character traits of users.

While virtue ethics offers a valuable perspective on the ethical implications of XR, it does have limitations. The reluctance of virtue ethics to impose rigid normative rules can be seen as both an advantage and a challenge. On the one hand, this flexibility allows for a more nuanced and context-sensitive ethical analysis, but on the other hand, it makes virtue ethics more difficult to apply in situations that require clear, actionable guidelines. Consequently, to provide a more robust and practical ethical framework, virtue ethics should be integrated with other normative frameworks, particularly those that offer concrete guidelines for ethical decision-making, such as **just consequentialism**.

While virtue ethics offers a valuable perspective on the ethical implications of XR, it does present challenges. One of the core characteristics of virtue ethics is its emphasis on personal character and moral development over time. However, this absence of explicit rules can make it difficult to apply to contexts that demand clear-cut ethical decisions, particularly when those decisions have immediate consequences. For instance, in the case of XR technologies used in distributed training, there may be situations where quick, actionable decisions are required, and virtue ethics alone does not provide a precise framework for determining the right course of action in those moments. Thus, while virtue ethics fosters long-term moral growth, it lacks the kind of actionable, rule-based guidelines that are often necessary in fast-paced, high-stakes environments. This gap suggests that virtue ethics should be used in conjunction with other ethical approaches—particularly those that offer more concrete criteria for evaluating and making ethical decisions.

Moor's [6] formulation of *just consequentialism* within the domain of computing ethics presents a noteworthy synthesis of deontological and consequentialist moral reasoning. By integrating principles derived from Rawls's [17] theory of justice with the evaluative strategies characteristic of consequentialism (e.g., Mill [18]), Moor proposes a normative framework tailored for ethical decision-making in technological contexts. This hybrid approach enables developers and practitioners to assess the ethical implications of their technological artefacts and practices through a dual lens: the intrinsic justifiability of actions and their foreseeable outcomes.

In this framework, an action is deemed ethically acceptable only if it satisfies two complementary conditions: first, that it can be morally justified with respect to established principles of justice and fairness (reflecting deontological considerations), and second, that it leads to outcomes which are beneficial or at least non-harmful (aligning with consequentialist criteria). This dual criterion ensures that ethical evaluation does not rest solely on abstract duties nor merely on utilitarian calculations but rather emerges from a balanced and context-sensitive interplay between the two traditions. In doing so, Moor's model offers a structured yet flexible ethical lens, especially suited to the complex and evolving challenges of information technology and software development. [6].

Heimo and Kimppa [7] argue that Moor's framework may be further enhanced by incorporating elements from virtue ethics. In particular, they suggest that the notion of *virtual virtue friendship* provides a meaningful avenue for integrating character-based ethical considerations into the structure of just consequentialism [7]. By doing so, they

introduce a dimension of moral development and personal character that complements Moor's emphasis on justifiable actions and favourable outcomes.

The inclusion of virtues, especially those cultivated within the context of digital and virtual interactions, enables a richer ethical evaluation—one that not only asks whether an action is just and effective, but also whether it contributes to the cultivation of moral character. In this view, the ethical assessment of technological practices extends beyond external consequences and procedural justifiability, encompassing also the internal dispositions and motivations of the agents involved. [7] Such an approach may help bridge the gap between abstract normative principles and the lived moral experience of individuals operating in increasingly complex technological environments.

4 Analysis

The inherited ethical challenges of distributed Extended Reality (XR), as a domain of XR experience, pedagogical solution and as an emerging technology, must be duly acknowledged. These challenges encompass, among others, safety, cybersecurity, accessibility, artificial intelligence, privacy, algorithmic bias, surveillance, automation, data ownership, cyber defence, environmental impact, technological discrimination, deep-fake technology, digital footprints, the future of work, transparency, AI-based decision-making, AI generated content, IPRs, technological unemployment, ethical usage, the dissemination of false information, and the manipulation of users (see e.g. [19–25]).

While these concerns form a significant part of the broader ethical landscape surrounding XR, their detailed examination falls outside the scope of the present study. Accordingly, the analysis presented here will focus primarily on the ways in which distributed XR differs from more traditional technological configurations (see Fig. 1). This places a requirement for understanding the ethical attributes and implications of different approaches.

One of the most significant changes introduced by distributed XR, in comparison to traditional, location-bound XR training, is the possibility for individuals to engage in training activities regardless of their physical location. While this increased flexibility offers considerable benefits, it also introduces certain ethical and practical concerns. Notably, users—particularly those with limited experience or knowledge—may inadvertently choose environments that are unsuitable or unsafe for training purposes. This may expose them, or those around them, to potential harm. Furthermore, the ability to select an appropriate and safe training environment is not always available to all users. A range of socio-economic and personal factors—such as living conditions, space constraints, or broader life circumstances—may significantly restrict access to optimal training settings. As a result, distributed XR systems must be evaluated not only for their technical functionality but also for the extent to which they account for these real-world disparities in users' ability to participate safely and effectively.

For users with limited technological proficiency or prior experience with XR environments, the absence of a physically present instructor may pose particular challenges, especially in relation to device operation and interaction within the virtual setting. This lack of immediate, in-person support to get the XR-equipment operational may hinder the effectiveness of training and, in some cases, lead to user frustration or disengagement.

Accordingly, sufficient emphasis must be placed on ensuring both the accessibility and usability of distributed XR systems. Design considerations may need to aim to accommodate a broad spectrum of users, possibly including those with minimal technical familiarity, to support equitable participation and meaningful learning outcomes across diverse user groups.

When compared to conventional forms of remote teaching and e-learning, the potential benefits of distributed XR may be substantial in certain domains. This is especially true in technical education, where the equipment involved is often expensive, large-scale, and logistically difficult to transport. In such cases, it is not uncommon for trainees to travel across countries or even continents to receive instruction. This type of training cannot be adequately delivered through standard video conferencing tools or traditional digital learning materials. In these contexts, remote teaching typically serves only a preliminary or supplementary function, lacking the capacity to provide meaningful, practice-based learning. Moreover, traditional e-learning methods often impose significant limitations on the range of learning techniques available to educators and students alike. Distributed XR, by contrast, facilitates ‘hands-on’ learning and learning-through-doing, offering immersive environments in which students can practise tasks, perform assessments, and even complete thesis-related work. For example, routine maintenance tasks on large industrial machines can be taught and examined virtually with a high degree of fidelity, reducing the need for physical access to equipment. Although distributed XR appears to offer a clear advantage in this comparison, it is not without limitations. Chief among these is the cost of the required hardware, which typically exceeds €500 per unit. This cost may pose a barrier to adoption, particularly in underfunded educational institutions or for learners in economically constrained situations.

Provided that appropriate safety measures are in place and sufficient attention has been given to the usability of the system, it may be observed—through a consequentialist lens—that the resource savings achieved, both fiscally and in terms of time, can outweigh the aforementioned challenges. Moreover, improvements in the efficiency of remote instruction may significantly enhance pedagogical effectiveness and enable the remote teaching of previously inaccessible fields, such as technical installation and maintenance. In this respect, the first condition of Just Consequentialism—namely, that the consequences of an action must be ethically favourable—can be regarded as fulfilled [6].

Nevertheless, the individual needs of learners and the requirements of specific user groups must not be overlooked. Efforts to maximise accessibility, enhance safety, and reduce systemic biases are essential. A practical example can be found in the spatial configuration of models in shared virtual environments: they must be positioned in ways that accommodate users of varying heights, given that physical characteristics such as stature are frequently projected into virtual representations. In the inherently hierarchical context of education, ethical sensitivity is required to mitigate the risk of replicating traditional structures of exclusion, such as racially biased camera systems [11].

Attention must also be paid to quality assurance in training content. Without rigorous oversight, there is a risk that distributed XR could become a lower-quality alternative disproportionately used by disadvantaged populations, while wealthier learners continue to benefit from high-quality in-person education. Still, it may reasonably be argued that

a world in which hands-on education is made accessible to a broader population—regardless of temporal or geographical constraints—is preferable to one in which such access remains limited to a privileged few.

From a virtue ethics perspective, focus must be directed toward how system design supports the cultivation of moral character. While distributed XR is primarily intended to foster knowledge and skill development, it should also be oriented toward supporting the formation of virtues and personal growth in its users. Among the virtues that such systems might help to nurture are *ambitiousness*, by providing high-quality learning tools; *magnificence*, by encouraging perseverance in challenging situations; *temperance*, through promoting goal-oriented self-regulation; and *courage*, in confronting novel tasks and striving to master them.

At the same time, developers and educators must also reflect on their own virtues in the process of designing and deploying such technologies. Qualities such as *truthfulness*, *kindness*, and *wittiness* play a crucial role in fostering supportive, meaningful, and human-centred learning environments. However, supporting the development of virtues—whether in oneself or in others—is rarely straightforward. There exists a constant risk of lapsing into *vanity*, *overambitiousness*, or *self-aggrandisement*. In such cases, the ethical integrity of the educational relationship may be compromised. The critical distinction lies in whether one seeks to support another's growth as their authentic self, or attempts instead to shape them into something else. If the aim is comparable to the mutual moral development found in virtuous friendships, then there is little room for ulterior motives or coercive dynamics. Supporting the moral development of others, particularly through technological systems, thus demands a high degree of wisdom and ethical discernment.

5 Conclusions

As an emerging technology, distributed XR is not devoid of ethically charged situations or underlying ethical challenges. It must be emphasised that the core problems of technology ethics, as well as the inherited challenges from digital learning and XR environments more broadly, remain relevant and applicable in the context of distributed XR solutions.

This study employed an extended version of Moor's theory, applying a multi-ethical analytical framework to evaluate the ethical viability of distributed XR systems. The framework functioned effectively in this context, enabling the analysis to reach the conclusion that, under certain conditions, distributed XR systems can be implemented in a manner that does not exacerbate existing ethical concerns.

Moreover, the applied method yielded a richer and more diverse interpretation than more linear or narrowly focused approaches typically allow, encouraging reflection on emerging issues from multiple ethical vantage points. Nevertheless, the complexity inherent in multi-ethical analysis introduces a methodological challenge of its own. The greater the analytical breadth, the higher the risk of overlooking essential details. This limitation, while characteristic of comprehensive approaches, must be acknowledged when interpreting the findings and applying them in practice.

Lastly, it should be noted that the present study does not engage in sufficient depth with the pedagogical specificities of distributed XR, nor with the particular ethical challenges that may emerge from these dimensions. As a result, future research will seek to

address these pedagogical aspects, applying the same ethical framework to generate more focused insights and to propose constructive contributions to responsible technological development.

Acknowledgments. This work is dedicated to the memory of James Moor (1942–2024), whose contributions to computer ethics continue to inspire.

References

1. Doolani, S., et al.: A review of extended reality (XR) technologies for manufacturing training. *Technologies* **8**(4), 77 (2020)
2. Lee, L.-H., Braud, D., Zhou, P., Wang, L., Xu, D., Lin, Z., Kumar, A.: All one needs to know about metaverse: a complete survey on technological singularity, virtual ecosystem, and research Agenda (2021). arXiv preprint [arXiv:2111.09795](https://arxiv.org/abs/2111.09795).
3. Fernandez, C.B., Hui, P.: Life, the metaverse and everything: an overview of privacy, ethics, and governance in the metaverse. In: Proc. 2022 IEEE 42nd Int. Conf. Distributed Computing Systems Workshops (ICDCSW), pp. 272–277 (2022)
4. Mystakidis, S.: Metaverse. *Encyclopedia* **2**(1), 486–497 (2022)
5. Immonen, T., et al.: Metaverse-based demonstrators as an alternative to traditional presentations: case fossil-free steelmaking processes. In: Proc. 16th Int. Conf. on Applied Human Factors and Ergonomics (AHFE 2025), Orlando, FL, July 26–30 (2025)
6. Moor, J.H.: Just consequentialism and computing. *Ethics Inf. Technol.* **1**(1), 61–65 (1999)
7. Heimo, O.I., Kimppa, K.K.: Virtuous just consequentialism: expanding the idea moor gave us. In: Societal Challenges in the Smart Society (pp. 35–40). Universidad de La Rioja.D. J. Solove, Understanding Privacy, Cambridge, MA: Harvard University Press 2008 (2020)
8. Hamad, A., Jia, B.: How virtual reality technology has changed our lives: an overview of the current and potential applications and limitations. *Int. J. Environ. Res. Public Health* **19**(18), 11278 (2022)
9. Meta: Meta quest 3: new mixed reality VR headset – Shop now | Meta Store. <https://www.meta.com/fin/en/quest/quest-3/>. Accessed: 8 May 2025
10. Apple: Apple vision pro–apple. <https://www.apple.com/apple-vision-pro/>. Accessed: 8 May 2025
11. Rajamäki, J., Tapola, M., Salonen, M., Heimo, O.I., Lehtonen, T.: Passthrough extended reality in maritime commissioning. In: Proc. 16th Int. Conf. on Applied Human Factors and Ergonomics (AHFE 2025), Orlando, FL 26–30 July 2025
12. Heimo, O.I., Kimppa, K.K., Helle, S., Korkalainen, T., Lehtonen, T.: Augmented reality—Towards an ethical fantasy?. In: Proc. 2014 IEEE Int. Symp. on Ethics in Science, Technology and Engineering, pp. 1–7 (2014)
13. Nyssönen, T., Heimo, O., Helle, S., Lehtonen, T., Mäkilä, T., Jauhiainen, J.: Anonymous collaboration in metaverse. In: Human-centered metaverse and digital environments. M. Zallio, ed. AHFE (2023) International Conference, AHFE Open Access, vol. 99, AHFE International, USA (2023). <https://doi.org/10.54941/ahfe1003937>
14. Aristotle: Nicomachean Ethics, ca. 350 BCE. Multiple translations used
15. McPherson, D.: Vocational virtue ethics: prospects for a virtue ethic approach to business. *J. Bus. Ethics* **116**(2), 283–296 (2013)
16. Vallor, S.: Social networking technology and the virtues. *Ethics Inf. Technol.* **12**, 157–170 (2010). <https://link.springer.com/content/pdf/10.1007%2Fs10676-009-9202-1.pdf>.
17. Rawls, J.: A theory of justice. Harvard University Press (1971)

18. Mill, J.S.: Utilitarianism (1861)
19. Floridi, L.: The Ethics of Information. Oxford University Press, Oxford (2013)
20. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: mapping the debate. *Big Data Soc.* (2016)
21. Koskinen, J.: Datenherrschaft—An ethically justified solution to the problem of ownership of patient information. *Finnish J. eHealth eWelfare* **8**(2–3), 139–142 (2016)
22. C. Wardle and H. Derakhshan, *Information Disorder: Toward an interdisciplinary framework for research and policy making*, Council of Europe report, 2017
23. Heimo, O.I.: Icarus, or the idea toward efficient, economical, and ethical acquirement of critical governmental information systems, Doctoral dissertation, University of Turku (2018)
24. Chesney, R., Citron, D.K.: Deep fakes: a looming challenge for privacy, democracy, and national security. *Calif. Law Rev.* **107**(6), 1753–1820 (2019)
25. Coeckelbergh, M.: AI Ethics, Cambridge. MIT Press, MA (2020)



Corporate Financial Statement Analysis in Education 4.0

Miguel Guillén-Pujadas¹ , David Alaminos¹ , Álvaro Carrasco-Aguilar² ,
and Mar Souto-Romero³ 

¹ Department of Business, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain
miguel.guillen@ub.edu

² Department of Marketing, Complutense University of Madrid. Av. Complutense,
s/n., Moncloa - Aravaca, 28040 Madrid, Spain

³ Department of Business Economics, University Rey Juan Carlos,
Paseo Artilleros, s/n., Vicálvaro, 28032 Madrid, Spain

Abstract. This article aims to analyze how the teaching of financial statement analysis can benefit from the use of artificial intelligence tools within the framework of Education 4.0. The integration of AI technologies can not only enhance students' technical understanding of financial analysis but also strengthen their analytical and critical thinking skills. The methodological design is based on a mixed approach that combines theoretical review with practical analysis of digital tools applied to financial education. The study proposes a comparison between a traditional approach to financial analysis, using spreadsheet for ratio calculations, and an AI-assisted approach through the ChatGPT Plus platform. AI enables the processing of large volumes of data, the identification of patterns, the forecasting of trends, and the evaluation of financial scenarios. The results show that incorporating AI into teaching improves the speed of calculation and the interpretation of financial data. However, solid training is required to avoid errors arising from automated use without conceptual understanding. The study highlights the need to reshape teaching methodologies by incorporating training in AI use as a key tool to prepare students for the challenges of a digitalized business environment. Furthermore, it suggests promoting future studies to empirically validate the impact of AI use on academic performance.

Keywords: Financial Statement Analysis · Artificial Intelligence · Education 4.0

1 Introduction

Teaching financial statement analysis is essential for training professionals capable of interpreting and making strategic decisions based on companies' economic situations. In an increasingly competitive and globalized business environment, having a solid foundation in this field enables students and future managers to understand the financial structure of organizations, identify their strengths and weaknesses, and assess their long-term viability. Teaching this subject not only provides technical knowledge but also

fosters analytical and critical skills, which are crucial for ensuring informed decision-making that supports business growth and sustainability (Monterrey & Sánchez, 2009; Peláez-Sánchez et al., 2024).

The use of artificial intelligence (AI)-based tools has significantly transformed financial statement analysis by enabling the processing of large volumes of data more rapidly. These technologies can facilitate pattern identification, trend forecasting, and financial scenario evaluation, thereby optimizing strategic decision-making (Kim, Muhn, & Nikolaev, 2024). However, their effectiveness depends on a deep understanding of their proper application, as improper use can lead to misinterpretations. In this context, integrating traditional teaching with AI aligns with the principles of Education 4.0, a model that promotes technology-based learning to develop digital, analytical, and critical skills (Armada Pacheco, 2023; Ruiz-Rojas et al., 2024). This combination allows students not only to acquire the theoretical foundations of financial analysis but also to strengthen their ability to accurately interpret and apply data in a highly digitalized business environment. The objective of this study is to compare the results proposed through traditional financial analysis with those generated using an artificial intelligence tool such as ChatGPT.

Through detailed analysis of financial reports, internal strengths and weaknesses can be identified, as well as external opportunities and threats, enabling strategic decision-making to ensure business viability. Financial statements reflect economic performance, debt management, and resource use efficiency, providing an objective basis for strategic planning and decision-making. To conduct effective financial analysis, it is essential to understand the methodology employed in preparing these documents. International regulations such as the International Financial Reporting Standards (IFRS) and Generally Accepted Accounting Principles (GAAP) ensure the consistency and comparability of information, guaranteeing its transparency and accuracy (Banco de España, 2024).

There are two main approaches to financial analysis: internal and external. The internal approach focuses on profitability, cost structure, and operational efficiency, using tools such as financial ratios, trend analysis, and liquidity assessment (Cordova, 2024). The external approach examines macroeconomic and market factors that may affect the company, such as competition, government policies, and global economic conditions. For this purpose, methods such as PESTEL analysis, Porter's Five Forces model, and benchmarking are employed (Piazza et al., 2025).

The article is organized as follows: first, the tools for internal and external financial analysis are presented. Section 3 describes the methodology and sample, while Sect. 4 presents the results obtained from both the traditional analysis and the AI-based analysis. Finally, Sect. 5 offers the conclusions and possible future research directions.

2 Tools for Conducting Financial Statement Analysis

The analysis of financial statements is a key tool for evaluating an organization's economic and financial health. To conduct this analysis effectively, it is essential to use a variety of tools that allow for examining both the internal aspects of the company and the external factors that may influence its performance (Beaver, 1966). Regarding internal analysis, financial statements offer a detailed view of the company's structure and

profitability through key documents such as the balance sheet and the income statement (Altman, 1968). On the other hand, external analysis considers factors such as the economic environment, industry dynamics, and benchmarking against other companies in the same sector, providing necessary context to evaluate the organization's competitive position and long-term viability (Basu, 1997). The combination of these tools allows for a comprehensive view of the company's financial and strategic situation, facilitating more informed and precise decision-making (DeFond & Zhang, 2014).

Artificial intelligence (AI) can be a valuable tool in both internal and external financial statement analysis. In internal analysis, AI facilitates the calculation and interpretation of accounting data, optimizing data processing flows and reducing the time required for these tasks (Kim, Muhn, & Nikolaev, 2024). In external analysis, AI can process macroeconomic and industry data from multiple sources, helping to identify external factors affecting the company's financial stability (Banco de España, 2024). However, it is crucial that AI models are fed with high-quality data and that their results are interpreted critically, as automation cannot fully replace human judgment in complex contexts (Cuervo, 2023).

The balance sheet allows for an internal examination of the composition of assets, liabilities, and equity, providing key information about solvency and financial stability (Kogan et al., 2009). Various indicators can be used for this purpose, such as the liquidity ratio, which measures the company's ability to meet its short-term obligations; the acid-test ratio, which offers a stricter evaluation of liquidity by excluding inventories and accounts receivable from current assets; the debt ratio, which reflects the degree of financial leverage; and the debt quality ratio, which analyzes the proportion of long-term debt relative to total debt (Monterrey & Sánchez, 2009).

The income statement, on the other hand, provides a detailed analysis of revenues and expenses, allowing for the evaluation of the organization's profitability and operational efficiency. Indicators such as Return on Equity (ROE), which measures the profitability of shareholders' equity and evaluates how efficiently profits are generated from shareholders' resources, and Return on Assets (ROA), which quantifies the profitability of the company's total assets, reflecting their capacity to generate income, can be employed (Li, 2010).

External analysis is crucial for understanding the environment in which a company operates and the potential impact of its strategies. It includes studying the economic environment, considering factors such as inflation, interest rates, and GDP growth; and the industry, including trends, regulations, and competitiveness, which allows for evaluating each company's relative performance (Brynjolfsson & McElheran, 2016). Tools such as SWOT analysis, the PESTEL model, and Porter's Five Forces can be used to examine competition and the factors affecting sector profitability (Piazza et al., 2025). These methodologies facilitate strategic decision-making and the identification of competitive advantages.

3 Sample and Methodology

With the objective of integrating traditional financial analysis with artificial intelligence-based tools in the educational field, a comparative study is proposed, applied to eight publicly traded construction companies in Spain: ACS S.A.; Duro Felguera S.A.; Obrascón Huarte Lain S.A.; Sacyr Construcción S.A.U.; Acciona Construcción S.A.; Ferrovial Construcción S.A.; FCC Construcción S.A.; and Grupo Empresarial San José S.A., covering the period from 2019 to 2023. The selection of this sample was based on the availability of complete and comparable data, and because it is representative of the publicly traded construction sector in Spain. This approach allows for evaluating the applicability of artificial intelligence in teaching, providing an innovative methodology for financial statement analysis. The integration of AI into financial education has been recognized as a powerful tool for enhancing learning and improving data analysis efficiency (Brynjolfsson & McElheran, 2016).

The study is conducted in three phases. In the first phase, the financial statements are analyzed using the traditional approach. In the second phase, the financial data of the companies are processed by ChatGPT Plus, which replicates the analyses applied in the initial phase. Finally, in the third phase, the results obtained through both approaches are compared, identifying the critical factors of each methodology. This exercise makes it possible to evaluate the effectiveness of artificial intelligence in financial analysis and its potential as a teaching tool in the classroom. The application of AI in this field has been shown to improve the predictive and analytical capabilities of financial statement analysis, optimizing risk assessment and financial performance evaluation processes (Kim, Muhn, & Nikolaev, 2024).

4 Results

4.1 First Phase

The first phase of the study focuses on the classical financial analysis of eight publicly traded Spanish construction companies: ACS, Ferrovial, Acciona, Sacyr, FCC, OHLA, Grupo San José, and Duro Felguera. In this stage, a spreadsheet is used without any type of automated tool or artificial intelligence. The entire process is conducted manually, with calculations performed directly by the analyst. The financial statements of each company for the period 2019–2023 are used. Based on this information, the main financial indicators are calculated in detail: liquidity ratios, debt ratios, debt quality ratios, return on assets (ROA), and return on equity (ROE). This approach allows for a deep understanding of each company's financial structure and the performance of its results over time.

The purpose of this phase is twofold: on one hand, to provide a rigorous exercise in analysis that reinforces the technical foundations of accounting analysis; and on the other, to generate a basis for comparison to evaluate, in later phases, the performance of artificial intelligence-based tools. In this way, a clear methodological starting point is established to assess the potential, limitations, and accuracy of alternative approaches supported by new technologies.

Table 1 shows the results obtained after calculating the main financial ratios for each of the eight companies using the traditional method. These indicators allow for a comparison of their financial situation and their evolution over time. In terms of liquidity, ACS stands out with a Current Ratio of 2.46, indicating a greater ability to meet its short-term obligations, while Duro Felguera and Sacyr present values below 1, reflecting potential difficulties in covering their current liabilities. Regarding debt levels, Duro Felguera has the highest Debt Ratio (1.40), suggesting a greater reliance on external financing, while ACS, with a Debt Ratio of 0.28, shows a lower level of financial leverage. In terms of debt quality, Sacyr presents the highest value (0.96), indicating that most of its debt is short-term, while Grupo San José shows the lowest level (0.28), suggesting a more balanced debt structure. In terms of profitability, ACS is the company with the highest ROA (7.45%) and ROE (14.52%), reflecting efficient management and a high return for shareholders. In contrast, Duro Felguera and OHLA present negative ROE values, indicating losses for their investors, while Acciona and Ferrovial record negative ROA values, reflecting low profitability on their assets.

Table 1. Financial ratios with spreadsheet

Company	Current R.	Quick R.	Debt R.	Debt Quality R.	ROA	ROE
ACCIONA	1.09	0.29	0.77	0.73	−0.45%	3.45%
ACS	2.46	1.08	0.28	0.26	7.45%	14.52%
DURO FALGUERA	0.74	0.21	1.40	0.84	5.46%	−16.13%
FCC	1.36	0.30	0.58	0.88	2.13%	25.05%
FERROVIAL	1.42	0.16	0.62	0.91	−0.20%	3.49%
GRUPO SAN JOSE	1.13	0.00	0.68	0.28	4.14%	8.71%
OHL	1.02	0.14	0.65	0.89	1.29%	−7.91%
SACYR	0.82	0.06	0.86	0.96	0.03%	−4.14%

4.2 Second Phase

In this second phase, the financial statements of the eight companies are processed using artificial intelligence tools such as ChatGPT Plus. The calculations and financial ratios performed in the initial phase using traditional methods are replicated. The objective is to assess AI's ability to automate and standardize financial analysis, facilitating the interpretation of complex data and improving the efficiency of the analytical process. The integration of ChatGPT into this process seeks to transform financial data tables into structured information quickly and accurately, providing an interactive experience that can enrich both teaching and professional practice in the financial field. By applying artificial intelligence to the financial analysis of these construction companies, this

study aims to demonstrate the feasibility and effectiveness of these tools as pedagogical and analytical assistants. This not only optimizes the time dedicated to data processing but also offers an innovative methodology that can be incorporated into educational and professional settings to enhance the understanding and application of key financial concepts.

The results obtained after calculations performed with ChatGPT are presented in Table 2. In terms of liquidity, ACS stands out with both a Current Ratio and Quick Ratio of 2.46, indicating a strong ability to meet short-term obligations, while Duro Felguera and Sacyr show values below 1, suggesting potential difficulties in covering their current liabilities without additional financing. In the analysis of indebtedness, Ferrovial has the highest Debt Ratio (1.66), indicating a strong reliance on external financing, followed by Duro Felguera (1.40). In contrast, ACS shows a Debt Ratio of 0.39, reflecting a lower level of financial leverage. Regarding debt quality, Ferrovial records the highest value (0.91), suggesting that most of its debt is short-term, while Sacyr presents the lowest value (0.04), indicating a greater proportion of long-term debt. In terms of profitability, FCC leads with a ROA of 10.48% and a ROE of 25.05%, reflecting efficient management and high returns for shareholders. ACS also shows strong figures with a ROA of 10.44% and a ROE of 14.52%. On the other hand, Duro Felguera, OHL, and Sacyr record negative ROA values, indicating losses on their assets, while OHL and Sacyr also report negative ROE values, reflecting negative returns for investors.

Table 2. Financial ratios with ChatGPT Plus

	Current R.	Quick R.	Debt R.	Debt Quality R.	ROA	ROE
ACCIONA	1.09	1.09	0.77	0.27	0.81%	3.45%
ACS	2.46	2.46	0.39	0.26	10.44%	14.52%
DURO FALGUERA	0.74	0.74	1.40	0.16	−6.53%	16.13%
FCC	1.36	1.36	0.58	0.12	10.48%	25.05%
FERROVIAL	1.42	1.42	1.66	0.91	1.31%	3.49%
GRUPO SAN JOSE	1.13	1.13	0.68	0.72	2.79%	8.71%
OHL	1.02	1.02	0.65	0.11	−2.77%	−7.91%
SACYR	0.82	0.82	0.86	0.04	−0.60%	−4.14%

As can be observed, there are differences between the two tables (1 and 2) due to the calculation methodologies used in each case, which can influence the interpretation of the financial results. These variations highlight the importance of clearly defining the criteria and assumptions employed in the analyses. In Phase 3 of the study, these differences will be further explored, assessing their impact and comparing manual approaches with automated tools to obtain a more accurate view of the financial performance of the companies analyzed.

4.3 Third Phase

In Phase 3 of this analysis, the results obtained through the traditional approach will be compared with those generated automatically through ChatGPT Plus for the eight selected companies: ACCIONA, ACS, DURO FELGUERA, FCC, FERROVIAL, GRUPO SAN JOSÉ, OHL, and SACYR. The study will focus on data from the years 2019 to 2023, with the aim of evaluating the effectiveness and accuracy of both approaches in the collection, processing, and analysis of information. For this purpose, the comparison has been sequenced by first taking the traditional value and subtracting the automated value. Thus, if the resulting figure is positive, the automated method would be underestimating the ratio, whereas if it is negative, it would be overestimating it. This approach allows for a more precise identification of the differences between the two methods and their impact on the interpretation of results.

The analysis of the differences between the traditional and automated methods across the eight selected companies reveals recurring patterns that characterize the behavior of the AI model concerning certain financial ratios. In general, the quick ratio shows a clear and consistent tendency to be overestimated by the automated approach across all companies, with negative differences typically ranging between -0.50 and -1.30. This is consistently observed in ACCIONA (Table 3), ACS (Table 4), DURO FELGUERA (Table 5), FCC (Table 6), FERROVIAL (Table 7), GRUPO SAN JOSÉ (Table 8), OHL (Table 9), and SACYR (Table 10), with an especially pronounced case in ACS during 2021 (-3.80). In contrast, the current ratio remains unchanged across all tables, with difference values of 0.00, suggesting greater stability in its calculation by the automated model. Regarding the debt ratio, the differences are generally low or negligible, although companies like FERROVIAL (Table 7) stand out by showing a sustained overestimation of indebtedness. As for the debt quality ratio, the results are more varied: while ACCIONA (Table 3), FCC (Table 6), OHL (Table 9), and SACYR (Table 10) show positive differences—suggesting that the model tends to underestimate this indicator—GRUPO SAN JOSÉ (Table 8) shows persistent negative values, reflecting a systematic overestimation.

Regarding profitability metrics, significant discrepancies are identified, particularly in ROA (Return on Assets), while ROE (Return on Equity) remains unchanged (0.00%) across almost all tables, except for DURO FELGUERA (Table 5), where extreme differences are observed, such as -227.94% in 2020. This case suggests that the automated approach may encounter difficulties when processing unusual or highly volatile financial scenarios. ROA, on the other hand, shows important variations over the analyzed years: for example, OHL (Table 9) shows a difference of 9.08% in 2019 and 4.06% in 2023, while DURO FELGUERA (Table 5) reaches a positive difference of 26.77% in 2020 and a negative difference of -9.36% in 2021, highlighting the model's sensitivity to drastic changes in operating profitability. Similarly, FCC (Table 6) presents a clear underestimation of ROA in 2023 (-8.35%) and FERROVIAL (Table 7) in 2022 (-4.75%). These discrepancies underline the need for a more thorough review when using the automated approach in contexts with volatile or atypical profitability.

In summary, although the use of ChatGPT helps to streamline the financial analysis process, the results demonstrate that its application still requires validation and adjustment, especially for indicators that depend on more complex interpretations or

are subject to external variability. Therefore, the importance of complementary analysis through traditional methods is reaffirmed, both to ensure accuracy and to provide the necessary context for financial decision-making.

Table 3. Acciona Differences

Year	Current R.	Quick R.	Debt R.	Debt Quality R.	ROA	ROE
2023	0.00	−0.80	0.00	0.46	−1.25%	0.00%
2022	0.00	−0.85	0.00	0.44	−0.62%	0.00%
2021	0.00	−0.83	0.00	0.57	−2.65%	0.00%
2020	0.00	−0.94	0.00	0.61	−2.65%	0.00%
2019	0.00	−1.09	0.00	0.56	−1.15%	0.00%

Table 4. ACS Differences

Year	Current R.	Quick R.	Debt R.	Debt Quality R.	ROA	ROE
2023	0.00	−1.38	−0.11	0.00	−2.99%	0.00%
2022	0.00	−1.08	−0.12	0.00	−1.16%	0.00%
2021	0.00	−3.80	−0.24	0.00	1.34%	0.00%
2020	0.00	−0.54	−0.62	0.00	−0.80%	0.00%
2019	0.00	−0.19	−0.48	0.00	−1.87%	0.00%

Table 5. Duro Falguera Differences

Year	Current R.	Quick R.	Debt R.	Debt Quality R.	ROA	ROE
2023	0.00	−0.52	0.00	0.69	11.98%	−32.27%
2022	0.00	−0.81	0.00	0.20	0.61%	0.00%
2021	0.00	−0.64	0.00	0.23	−9.36%	0.00%
2020	0.00	−0.54	0.00	0.66	26.77%	−227.94%
2019	0.00	−0.65	0.00	0.80	−1.48%	0.00%

Table 6. FCC Construcción Differences

Year	Current R.	Quick R.	Debt R.	Debt Quality R.	ROA	ROE
2023	0.00	−1.05	0.00	0.77	−8.35%	0.00%
2022	0.00	−0.83	0.00	0.79	2.05%	0.00%

(continued)

Table 6. (continued)

Year	Current R.	Quick R.	Debt R.	Debt Quality R.	ROA	ROE
2021	0.00	−1.03	0.00	0.74	−6.98%	0.00%
2020	0.00	−0.91	0.00	0.70	2.95%	0.00%
2019	0.00	−0.70	0.00	0.73	3.25%	0.00%

Table 7. Ferrovia Differences

Year	Current R.	Quick R.	Debt R.	Debt Quality R.	ROA	ROE
2023	0.00	−1.26	−1.04	0.00	−1.51%	0.00%
2022	0.00	−1.29	−0.63	0.00	−4.75%	0.00%
2021	0.00	−1.28	−0.69	0.00	−2.62%	0.00%
2020	0.00	−1.27	−0.84	0.00	2.12%	0.00%
2019	0.00	−1.27	−0.70	0.00	−1.30%	0.00%

Table 8. Grupo San José Differences

Year	Current R.	Quick R.	Debt R.	Debt Quality R.	ROA	ROE
2023	0.00	−1.13	0.00	−0.44	1.35%	0.00%
2022	0.00	−1.09	0.00	−0.50	1.88%	0.00%
2021	0.00	−1.08	0.00	−0.42	1.74%	0.00%
2020	0.00	−1.24	0.00	−0.48	6.43%	0.00%
2019	0.00	−0.60	0.00	−0.32	−10.70%	0.00%

Table 9. OHL Differences

Year	Current R.	Quick R.	Debt R.	Debt Quality R.	ROA	ROE
2023	0.00	−0.89	0.00	0.79	4.06%	0.00%
2022	0.00	−0.77	0.00	0.30	3.77%	0.00%
2021	0.00	−0.72	0.00	0.36	−1.55%	0.00%
2020	0.00	−0.69	0.00	0.32	5.93%	0.00%
2019	0.00	−1.45	0.00	0.33	9.08%	0.00%

Table 10. Sacyr Differences

Year	Current R.	Quick R.	Debt R.	Debt Quality R.	ROA	ROE
2023	0.00	−0.76	0.00	0.91	0.63%	0.00%
2022	0.00	−0.75	0.00	0.91	2.90%	0.00%
2021	0.00	−0.79	0.00	0.82	1.93%	0.00%
2020	0.00	−0.82	0.00	0.83	−6.63%	0.00%
2019	0.00	−0.91	0.00	0.82	−0.92%	0.00%

The implementation of artificial intelligence tools such as ChatGPT in the study of financial analysis proves especially valuable within the Education 4.0 model, as it not only allows for the automation of routine processes like the calculation of financial ratios but also provides students with real opportunities to develop key competencies such as critical data validation, contextual interpretation of results, and analytical thinking. As evidenced in the comparative analysis of the results across the eight companies (Tables 3, 4, 5, 6, 7, 8, 9, and 10), the automated approach successfully replicates certain indicators —such as the current ratio or ROE— but also reveals significant differences in more sensitive metrics like the quick ratio or ROA, particularly in contexts of high financial volatility. These discrepancies, far from being limitations, can become powerful pedagogical resources: they allow students to identify margins of error, reflect on the causes behind the differences, and strengthen their professional judgment by contrasting methods. In this regard, the use of AI in the classroom fully aligns with the principles of Education 4.0, as it promotes a more active, autonomous, and situated learning process, where technology does not replace human judgment but rather enhances it by placing the student at the center of financial analysis and decision-making.

5 Conclusions and Limitations

Education 4.0 promotes the integration of technologies such as artificial intelligence (AI) into teaching processes, with the goal of developing students who are more autonomous, critical, and digitally competent (Kim, Muhn, & Nikolaev, 2024; Peláez-Sánchez et al., 2024; Ruiz-Rojas et al., 2024). This study proposes analyzing the viability and relevance of using an AI tool in the development of financial statement analysis, comparing its performance against the traditional approach based on manual calculations and consultations of primary sources. The main objective was to evaluate to what extent automation through AI can replicate, complement, or enhance financial analysis in educational environments, by calculating key ratios for eight companies in the construction and infrastructure sector between 2019 and 2023.

The study involved applying both traditional and automated approaches to the calculation of six financial ratios: current ratio, quick ratio, debt ratio, debt quality ratio, ROA, and ROE. Subsequently, the difference between both methods was estimated, using the traditional approach as a reference, to identify possible deviations. The results revealed that artificial intelligence correctly reproduces constant values such as the debt ratio or

ROE (with differences close to zero), but presents significant variations in other indicators, particularly in the quick ratio and ROA. These discrepancies remained relatively consistent across most of the companies analyzed. While AI proves effective for certain calculations, it still requires adjustments or supervision when dealing with complex data or ratios sensitive to small variations in liquid assets or net results.

It is evident that the use of AI in the teaching of financial analysis holds high educational potential, as it allows for the automation of processes, time savings, and the development of students' critical interpretation and result-contrast skills, fully aligned with the principles of Education 4.0. Nevertheless, limitations related to the reliability of the data generated and the lack of precision in specific financial contexts were identified, reinforcing the need to complement these tools with rigorous training in accounting and financial fundamentals. Potential future lines of research proposed include exploring the impact of sustained AI use on the development of analytical competencies in economics and business administration students, as well as integrating intelligent platforms into formative assessment systems within virtual learning environments.

Acknowledgments. This research was supported by Telefonica and the Telefonica Chair on Smart Cities of the Universitat de Barcelona and Universitat Rovira i Virgili (project number 42.DB.00.18.00).

References

- Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **23**(4), 589–609 (1968)
- Armada Pacheco, J.M.: Challenges of university teaching in the face of Education 4.0. *e-Revista Multidisciplinaria Del Saber* **1**, e-RMS01052023 (2023)
- de España, B.: Artificial intelligence in the financial system. *Financ. Stab. Rev.* **47**, 123–145 (2024)
- Basu, S.: The conservatism principle and the asymmetric timeliness of earnings. *J. Account. Econ.* **24**(1), 3–37 (1997)
- Beaver, W.H.: Financial ratios as predictors of failure. *J. Account. Res.* **4**, 71–111 (1966)
- Brynjolfsson, E., McElheran, K.: The rapid adoption of data-driven decision-making. *Am. Econ. Rev.* **106**(5), 133–139 (2016)
- Cordova, M.: The benefits of Artificial Intelligence in the development of accounting activities. *Revista de Contabilidad y Finanzas* **15**(3), 45–60 (2024)
- Cuervo, R.: Predictive AI for SME and large enterprise financial performance management (2023). *arXiv preprint* [arXiv:2311.05840](https://arxiv.org/abs/2311.05840)
- DeFond, M.L., Zhang, J.: A review of archival auditing research. *J. Account. Econ.* **58**(2–3), 275–326 (2014)
- Kim, A., Muhn, M., Nikolaev, V.: Financial statement analysis with large language models (2024). *arXiv preprint* [arXiv:2407.17866](https://arxiv.org/abs/2407.17866)
- Kogan, S., Levin, D., Routledge, B.R., Sagi, J.S., Smith, N.A.: Predicting risk from financial reports with regression. *J. Financ.* **64**(1), 235–265 (2009)
- Li, F.: The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *J. Account. Res.* **48**(5), 1049–1102 (2010)
- Monterrey, J., Sánchez, L.: A teaching strategy for financial statement analysis. *Revista de Contabilidad* **12**(2), 285–312 (2009)

- Piazza, M., Passacantando, M., Magli, F., Doni, F., Amaduzzi, A., Messina, E.: Explainable Artificial Intelligence for identifying profitability predictors in financial statements (2025). arXiv preprint [arXiv:2501.17676](https://arxiv.org/abs/2501.17676)
- Peláez-Sánchez, I.C., Velarde-Camaqui, D., Glasserman-Morales, L.D.: The impact of large language models on higher education: exploring the connection between AI and Education 4.0. *Front. Educ.* **9**, 1392091 (2024)
- Ruiz-Rojas, L.I., Salvador-Ullauri, L., Acosta-Vargas, P.: Collaborative working and critical thinking: adoption of generative artificial intelligence tools in higher education. *Sustainability* **16**(13), 5367 (2024)



Challenging AI as Critical Thinking

Michael S. Kirkpatrick^(✉) 

James Madison University, Harrisonburg, VA 22807, USA
kirkpams@jmu.edu

Abstract. Generative AI systems, such as LLMs, have rapidly become popular tools for students and professionals alike. While the results of these systems can often be useful due to their underlying statistical properties, these tools can produce erroneous and harmful results if not evaluated critically. In this work, we discuss the nature of LLMs, their potential benefits for education, and their flaws. In particular, we highlight both epistemological and moral concerns that users should heed when using them. To address these concerns, instructors should model how to interact with these tools as a form of critical thinking. Our goal in this article is to provide the necessary context and potential directions for using these tools successfully to augment the social process of education.

Keywords: active learning · artificial intelligence · critical thinking · education

1 Introduction

“Education is a social process; education is growth; education is not preparation for life but is life itself.” This quote by John Dewey appears in his book *The Social and the Society*. Or, at least, that is what the AI overview provided by Google declared in response to a query about the origin of this quote. One problem with this summary is that there is no record of this book’s existence¹. This exchange illustrates the truth of Dewey’s claim. Education is inherently a social process built on relationships, communication, empathy, and many other social skills. In particular, throughout the education process, students rely on instructors with sufficient expertise who act in good faith.

Recent advances in the field of artificial intelligence (AI), specifically the creation of generative AI tools, threaten to disrupt this dynamic in many ways. In particular, large language models (LLMs) provide a conversational interface that allows students to interact with AI to explore knowledge in new ways. On the one hand, students can chat with the LLM, asking it to paraphrase a complex piece of text, requesting additional problems to work through, or seeking possible sources and citations for a research project. On the other hand, students can also use LLMs to cheat, asking the tools to generate answers for a homework assignment or entire essays.

¹ One of Dewey’s most famous works has the similar title *The School and Society* (1899). However, the quotation does not appear in this text.

Regardless of whether students are using LLMs for benevolent purposes, the introduction of the tools raises a number of issues that educators must confront. Initial concerns have focused on the accuracy of these tools; using LLMs to search for knowledge can send students and other users toward incorrect or even malicious results. The problem goes deeper than just the question of correctness, however, as LLMs are based on probabilistic computational architectures that are built on correlations rather than rules of inference and domain expertise. That is, unlike an instructor, LLMs cannot be relied on as possessing sufficient expertise or acting in good faith; there is an inherent and unavoidable difference between trust placed in an LLM as opposed to an instructor.

As LLMs and other AI tools offer great promise in many areas, it is imperative that students learn to use them responsibly. To do so, students must learn to recognize the epistemological shortcomings of the AI models and to challenge the outputs critically. Instructors play a key role in this training, as they can leverage the social relationships underlying the traditional classroom model to demonstrate how to use LLMs appropriately. That is, instructors must help students to recognize that challenging LLM outputs is a fundamental critical thinking skill.

In this article, we will discuss how LLMs work and how they can be used to support the learning process. We will then examine the shortcomings of these systems that emerge from their probabilistic and computational nature. Specifically, we will discuss the problems of “hallucinations” and how stronger critiques of these systems emphasize the underlying epistemological flaws of LLMs. We will also highlight the lack of moral agency of LLMs, contrasting them with the good faith assumptions underlying the student-instructor relationship. We will then close with a discussion of how instructors could integrate LLMs into their courses to model how to use these tools effectively and appropriately.

Ultimately, AI tools—and LLMs, in particular—are popular tools and will continue to gain widespread use in education. Learning to challenge AI systems has become an essential part of the critical thinking learning outcomes that lie at the heart of education. Our goal in this article is to provide the necessary context and potential directions for using these tools successfully to augment the social process of education.

2 The Structure and Functioning of LLMs

Large-language models (LLMs) are a category of AI models that use deep learning on a large set of input data to process and respond to natural language requests. LLMs can be built on a variety of architectures, including recurrent neural networks (RNN), convolutional neural networks (CNN), or transformers. Generative pretrained transformers (GPT) is one type of LLM that is particularly effective for processing users’ requests and generating new text as a response [14]. This transformer architecture is the underlying structure of popular chat-based LLMs, including OpenAI’s ChatGPT, Anthropic’s Claude, and Google’s Gemini.

The key to understanding these architectures is the recognition that they are all probabilistic in nature; these models are trained by providing a large corpus of input data, and the model itself incorporates a probability distribution of the frequency of words and their correlations. The new text in response to a query has a high probability

of responding correctly to queries due to the statistical correlations the GPT derived from the training data. Neural networks and transformers simply do not incorporate rules of logic or reasoning; their ontologies are derived from statistical association of words.

To illustrate the functioning of these tools, we returned to the quote from opening paragraph of the introduction. We provided the following prompt to three popular chat-based LLM tools:

What is the source of the John Dewey quote, “Education is a social process; education is growth; education is not preparation for life but is life itself.”?

When we posed this question to ChatGPT (GPT-4o), the tool indicated that the quote “is widely attributed to John Dewey and encapsulates his educational philosophy. While this exact phrasing does not appear verbatim in his published works, it synthesizes key themes from his writings.” In addition, ChatGPT identified² a similar (and real) quote “I believe that education, therefore, is a process of living and not a preparation for future living.” in his 1897 article “My Pedagogic Creed,” while noting similar themes in his 1916 book *Democracy and Education*. ChatGPT ends by noting that “while the specific quote in question may be a paraphrase, it accurately reflects Dewey’s perspective that education is an integral part of life and personal development, rather than merely a preparation for future endeavors.”

Gemini (2.0 Flash) provides a very similar response, nothing that “while this specific phrasing is very closely aligned with Dewey’s educational philosophy, it is more accurately considered a **paraphrase** [*emph. Orig.*] of ideas he expressed in several of his works, rather than a direct, verbatim quotation.” Gemini also identifies the same two works as ChatGPT, and states that the quote is “best understood as a concise summary of his core beliefs about the nature and purpose of education, drawn from the broader context of his writings.” Gemini did not identify the similar quote that ChatGPT found.

In contrast to the other tools, Claude (3.7 Sonnet) declared that the quote in the prompt “comes from John Dewey’s book ‘Democracy and Education,’ published in 1916.” Claude additionally states that, in the book, “Dewey argues against the view that education is merely preparation for future life, instead emphasizing that education is an ongoing process embedded in lived experience.”

These results highlight a number of important facets of working with LLMs. First, the accuracy of the results and the additional context provided can vary from tool to tool. While ChatGPT and Gemini both correctly identified that the quote is a paraphrase of Dewey’s ideas (though often attributed to Dewey as a quote), Claude incorrectly states that the quote appears in *Democracy and Education*.

Furthermore, Claude’s response is troubling in that it could be read to suggest that “education is [not] merely preparation for future life” but is rather “an ongoing process embedded in lived experience” is a primary theme of the book. In fact, asking the follow-up question, “What is the main theme of John Dewey’s *Democracy and Education*?”

² One challenge of writing about LLMs is to avoid anthropomorphizing these systems through word choices. Phrasing such as “ChatGPT identified” suggests an agency and intentionality that should not go unexamined. While we will examine this aspect later in the paper, we find that such phrasing is convenient shorthand for more verbose (and less readable) sentiments such as “the output from the ChatGPT tool contained a reference to...”.

elicits a list of five themes, the first of which is, “Dewey rejects the notion that education is merely preparation for future life; instead, it’s an ongoing social process embedded in present experience.” In contrast, ChatGPT more accurately states that “the relationship between education and democracy, with the central idea that education is both a social and individual process necessary for the functioning of a democratic society,” while Gemini states that the main theme is “the indispensable connection between education and a functioning democratic society.” Claude’s response—which focuses on the individual rather than the societal impact—seems to echo the original prompt’s phrasing more than providing a true summary of the work. This distinction could be easily missed, leading the user to a poorer understanding of Dewey’s work.

Astute readers may also have detected the divergence between these responses and the response in the introduction. This divergence is the result of a three-month time difference in the tools. The introduction’s response was based on a query to the main Google search engine in January 2025, shortly before the release of the updated Gemini 2.0 Flash; the more recent queries were run using the specific tools in April 2025. Consequently, these tools evolve rapidly, which makes their responses very ephemeral. Furthermore, even when the tools are updated toward more accurate responses, there is no notice to the user that earlier responses were incorrect. Explicit feedback and correction are an important part of the learning process that may be missed through the use of these tools.

These models are also used as the basis for more specialized purposes, such as the code generation (GitHub Copilot, Cursor), image creation (DALL-E), and educational tools (Duolingo, Khan Academy). To illustrate the possible range of outputs, we provided the following prompt to Claude and ChatGPT (Gemini does not currently support image generation):

Create an image of John Dewey chatting with Leonard Bernstein.

When using both tools in a conversational manner, the initial responses described what the intended images would contain. The response from Claude indicated that “Dewey is shown on the left with his characteristic beard and glasses, representing his scholarly appearance. Bernstein is on the right with his distinctive hairstyle and trademark bow tie, which he often wore while conducting.” In the image, a word bubble near Dewey states, “Education is life itself!” while the one near Bernstein states, “And music is the universal language of connection!” Claude goes on to state that “They’re portrayed in a cozy setting with bookshelves in the background (representing Dewey’s philosophical works) and a piano (representing Bernstein’s musical world), having what appears to be a thoughtful exchange about education and music as transformative experiences.”

ChatGPT’s response indicated, “That sounds like a fascinating scene—John Dewey (the philosopher) chatting with Leonard Bernstein (the composer and conductor)! Since you requested an image, I’ll create a **generated artistic image** [*emph. Orig.*] showing them in conversation—perhaps seated, Dewey looking thoughtful and Bernstein animated, maybe even near a piano.” The images generated are shown below.



The output from Claude (left) and ChatGPT (right) when asked to generate an image of John Dewey chatting with Leonard Bernstein.



Both of these initial responses are obviously fabricated. While the Claude result differs significantly from what the tool describes, the ChatGPT result does a much better job of creating an artistic depiction. However, ChatGPT also provides a more direct “Create image” button that does not provide the same conversational output. Instead, the tool provides a message that the image is being generated. Using this button with the text “John Dewey chatting with Leonard Bernstein” produced the image shown to the right. Although the positioning of the two men is somewhat awkward and close together (their arms seem to merge near the bottom), the image is significantly more plausible than the artistic image. Additional context that Bernstein was only 34 years old when Dewey died at the age of 92, however, makes it clear the image as generated.

3 LLMs in Education and Beyond

In many circumstances, LLMs are very powerful and effective tools that can contribute to the field of education. LLMs have been used to help non-English speakers engage with communities where English is the dominant language [26]. In education, this is particularly important, as reducing language barriers for students makes learning more accessible [28]; students have successfully used these tools to improve necessary skills, such as essay writing and note taking. Using LLMs for this purpose can help to broaden participation in a variety of contexts.

Even when language is not ostensibly a barrier, different communities can fail to communicate successfully. These types of communication barriers can exist between experts and the general population, and LLMs can help to bridge this gap. Within the medical community, doctors have used ChatGPT to improve their communication with patients [18]. Such practices could be incorporated into professional training in medicine and a variety of other fields, improving professional communication with those who lack subject matter expertise.

Responsible use of LLMs in education should be based on practices that have been demonstrated to be most successful for learning outcomes. Research within the education community has long advocated for a transition to active learning. Active learning has been shown to increase learning outcomes significantly in a variety of settings when compared to traditional lecture-based instruction [10]. One common way to summarize the resulting recommendations for learner-centered teaching [31] is to describe a shift from instructors as a “sage on the stage” to a “guide on the side” [16]. As a part of this shift, instructors must explicitly teach and model many necessary skills, such as collaboration [23], that are not specific to their domain of expertise.

One interesting approach using such a learner-centered model demonstrates how to use LLMs within the classroom for building students’ knowledge bases [6]. ChatGPT was used in two different ways to explore topics related to religious literacy. In one phase, the teacher used ChatGPT to create a sample essay as a model of “an essay that is well-written but fails to use in-text citations correctly.” After critically analyzing the essay, students integrated the generated essay’s ideas with their own for new research directions. The tool was also used to generate questions to guide their study. In a second phase, students took more of the lead, using ChatGPT to generate research ideas and used ChatGPT to respond to their own questions. That is, the teacher initially led the students through exercises of how to use LLMs for knowledge building as scaffolding for their own use of the tools.

4 Hallucinations, Stochastic Parrots, or Bullshit Machines?

Although LLMs have many potential benefits, the widespread adoption of such tools has raised a number of concerns and questions. For instance, De Angelis et al. [8] highlighted concerns that LLMs can threaten public health by spreading misinformation. LLMs may generate text responses that appear plausible but actually produce empirically false claims, as the earlier example of the Dewey quote illustrates. The debate over the nature of such systems has a long history [29]. In more modern terms, LLMs are subject to the problem of hallucination [27, 30]. To illustrate the concerns of hallucinations, we provided the following prompt relating Dewey and Bernstein to Claude, ChatGPT, and Gemini:

Create a reading list for learning about how John Dewey influenced Leonard Bernstein.

We are not aware of any particular scholarly works describing such an influence. The responses from both Claude and Gemini began with a similar result. Claude indicated that “there isn’t extensive documentation of direct influence between John Dewey and

Leonard Bernstein,” while Gemini stated that “a direct, comprehensive study on John Dewey’s influence on Leonard Bernstein might be limited.” However, both tools proceeded to create a reading list on Dewey’s influence regardless of this disclaimer. Claude stated that “their educational philosophies share remarkable parallels. Both emphasized experiential learning, democratic access to knowledge, and education as an active, participatory process.” Gemini suggested “we can create a reading list by exploring their individual philosophies and identifying potential connections.”

Analyzing the list of works included in the results highlights potential concerns of using LLMs for this kind of work. Both Claude and Gemini identified Dewey’s books *Art as Experience* (1934) and *Democracy and Education* (1916) as relevant. However, Claude identified Bernstein’s *The Unanswered Question: Six Talks at Harvard* (1976) and *The Joy of Music* (1959) as potential sources, while Gemini suggested Bernstein’s book *Findings* (1991) and his televised *Young People’s Concerts* (1959 - 1972). These results could provide a plausible starting point for analysis, as they all involve questions of education and aesthetics. However, Claude’s list also includes the following citations:

1. Freedland, M. (2018). *Leonard Bernstein: The political life of an American musician*. University of California Press.
2. Burton, H. (1994). *Leonard Bernstein: A life*. Doubleday.
3. Bowie, A. (2009). “Music, philosophy, and modernity.” *Cambridge University Press*, 248–267.
4. Johnston, J.S. (2006). “Inquiry and education: John Dewey and the quest for democracy.” *SUNY Press*, 89–112.
5. Talebi, K. (2015). “Art and democracy: The pedagogical ideas of Leonard Bernstein and John Dewey.” *Journal of Aesthetic Education*, 49(1), 1–19.

All of these items are incorrect. In this list, item 1 (Freedland, 2018) refers to a 2009 book by Barry Seldes. Item 2 (Burton, 1994) merges two books: while Humphrey Burton’s *Leonard Bernstein* was published in 1994, the book *Leonard Bernstein: A Life* was written by Meryle Secrest and published in the same year. Items 3 and 4 are both books though they are formatted as articles. Item 5 is the most troubling, as this article does not exist (though the journal itself does).

Unlike Claude and Gemini, ChatGPT responds to the original prompt confidently asserting, “Absolutely—here’s a **focused reading list** on how **John Dewey’s philosophy** influenced **Leonard Bernstein’s** work, especially Bernstein’s educational and artistic philosophy [*emph. Orig.*]” The reading list includes similar works returned by Claude and Gemini, while also including the following fabricated article (claiming that it can be “found in journals like *Philosophy of Music Education Review*”):

6. DeLorenzo, L.C. (2012). “A Musical Morality: Leonard Bernstein’s Educational Legacy and the Philosophy of John Dewey.”

Although the problem of hallucinations is frequently characterized as an issue of accuracy, the problem is more fundamentally an issue of the epistemology of these systems. As such, more recent critiques challenge the claim that these systems actually exhibit features that should be considered “intelligent.” Since LLMs are based on statistical correlations between words and they do not build on principles of reasoning,

experimentation, or domain expertise, they have been characterized as “stochastic parrots” [2] rather than intelligent agents. Recall from the original Dewey quote example that Claude’s response suggested the most important theme of *Democracy and Education* was the point that “Dewey rejects the notion that education is merely preparation for future life; instead, it’s an ongoing social process embedded in present experience.” As we noted previously, this response seems to be echoing back the original quote in our query, suggesting that Claude’s response fits the “stochastic parrot” framing.

More recent critiques of LLMs make an even stronger claim. In particular, some are now arguing that it is best to view their results as “bullshit” in the sense described by Frankfurt [9]. In this sense, the generated output of LLMs should be treated as being made with “indifference towards the truth of the utterance” [12] or “intended to appear authoritative or persuasive without regard to its actual truth or logical consistency” [3]. Recall ChatGPT’s confident assertion that there were several works analyzing Dewey’s influence on Bernstein. Although it is plausible that there may be a connection, none of the sources identified by any of the tools explicitly address both Dewey and Bernstein. As such, their responses—particularly ChatGPT’s confident assertion—fail to evaluate the veracity of the claimed connection beyond correlating similar terms in secondary sources about these works, demonstrating an indifference toward the underlying truth. Given that the output from these systems may differ sharply from the knowledge shared by domain experts, the “bullshit machine” label may be warranted, given the lack of moral agency of the systems.

5 LLMs and Moral Agency

A core aspect of the dilemma regarding moral agency and critical thinking in relation to LLMs is that these systems exhibit a large-scale version of the problem of many hands. Nissenbaum [22] identified this problem as a barrier for accountability while arguing for those who create computer system to adopt a stronger standard of care. Blaming the system for its inaccurate results is making the computer into a scapegoat, inappropriately assigning the moral agency to the system rather than those who build the tools and the models. Similarly, Johnson [15] has argued that computer systems such as LLMs are not moral agents that can accept responsibility for claims of truth and trustworthiness.

Pettit [25] described a framework for thinking about group agency that is also relevant to the discussion of LLMs. This model identifies five elements for assigning group agency. Two of these elements are an “authorized voice” and a “constructed voice,” which together describe the nature of LLMs. However, the LLM itself cannot demonstrate a “capacity for commitment” to establish and maintain the veracity of their results; the ephemeral nature of software updates and versions obviate this. Similarly, LLMs and other AI tools exhibit an opaque nature that makes introspection to the rationale for a response impossible to determine. As such, it cannot be said that the system designers are acting with any intention to affirm the results. In short, LLMs simply are not capable of being established as moral agents. Adding another layer of indirection to move from experts to novices with insufficient experience exacerbates this concern.

The field of radiology provides important insights regarding the implications of widespread use of and reliance on AI tools such as LLMs. In comments to a machine

learning conference in 2016, Geoff Hinton famously predicted that radiologists would become obsolete given advances in generative adversarial networks (GANs) [13]. As of 2024, though, the field faced a labor shortage [5]. Rather than replacing radiologists with AI systems, there has been a shift to exploring the use of AI predictions to support radiologists in their work. Unfortunately, there is growing evidence that this approach may lead to worse outcomes. In one recent study, the performance of radiologists *decreased* when they were provided with predictions generated by AI [1]. In that study, the largest performance drops occurred in cases where radiologists were less certain of their initial determination; rather than relying on their own expertise and judgment, they were then more likely to defer to the AI predictions, which were more commonly wrong. In short, strongly hyped claims about AI has discouraged people to pursue the field and is likely to lead to worse outcomes than before.

The moral concerns regarding LLMs and other AI tools go beyond labor shortages and inaccuracies³. Gendron et al. [11] expressed concern that widespread use of AI tools could deskill and degrade core activities in academic publishing, a loss similar to the labor shortage in radiology. Yang et al. [32] have demonstrated that AI tools for medical imaging can infer demographic characteristics and produce biased results in cases where there is no basis for using this information in medical diagnoses. The concern that users become overly reliant on AI judgments adds to these fairness problems, as there is growing evidence that police departments unquestioningly use facial recognition systems when making arrests [20]. Similar issues of bias and fairness have been demonstrated specifically in LLMs [17]. Even more troubling, many of these approaches fail to account for the threats created by intentional manipulations of data sets to create malicious models [7].

Taken together, the fact that LLMs lack moral agency raises many concerns. These results have shown that overly deferential users—even if they have subject-matter expertise—can allow harm that goes beyond simple inaccuracies. The complex nature of these phenomena—including labor shortages, emergent biases, and misdiagnoses—will render accountability extremely difficult. Human educators and experts inherently have more agency and provide significantly stronger—though by no means perfect—assurances for accountability.

6 Toward AI Literacy as Critical Thinking

Up to this point, we have examined several key elements of the nature of LLMs. These tools provide a conversational interface to respond to users' queries based on probabilistic architectures that model the correlations between words in a large corpus of training data. The massive scale of these models facilitates the construction of plausible query responses that reflect the collective knowledge that exists in that training set. However, the models have no built-in constructs of ground truth, logic, or reasoning, suggesting these

³ Readers who are well-versed in issues around LLMs will note that our list of concerns omits many important issues, such as the neglect of intellectual property rights and environmental cost of building the models. While these are valid concerns, we have narrowed our focus here to issues most relevant to the topic of critical thinking.

constructed responses are intended to appear authoritative despite being indifferent to their veracity; furthermore, these models lack moral agency for the problems they cause.

Taking a balanced view of LLMs in the context of education requires considering both the potential merits and the flaws of such systems. Additionally, their popularity and widespread adoption makes it unlikely that these systems can be made to simply disappear if their flaws cannot be addressed. Instead, these tools necessitate another shift for instructors toward explicit instruction and modeling. There must be greater recognition and emphasis on the notion that AI output should be treated as Frankfurtian “bullshit.” Questioning and challenging AI output has become a vital critical thinking skill and instructors need to shift their pedagogy to model and cultivate this behavior.

The value of critical thinking as a skill and methods for teaching it have long been studied. For instance, Murawski [21] described critical thinking as a method for improving one’s thinking; the goal of the process is “to determine strengths and weaknesses in one’s thinking in order to maintain the strengths and make improvements by targeting the weaknesses.” Furthermore, these are skills that can be taught and they “are transferable from the classroom to the workplace,” which is fundamental to the functioning of society. Brookfield [4] also argued that critical thinking is a social process, emphasizing the role of both peers and instructors in the process.

Given the moral concerns regarding LLMs, along with their widespread popularity, it is incumbent on instructors to take an active role in teaching students how to approach these tools critically. Additionally, it is important to recognize that much of this work needs to be done in an active learning setting rather than as independent assignments; LLMs are so easy to use that students would have a strong incentive to use them to cheat [19], thus defeating the purpose of the assignment. Here, we sketch some possible approaches based on several critical thinking strategies identified by Paul and Binker [24].

Evaluating the credibility of sources of information. When we used the LLMs to generate a reading list to link Dewey and Bernstein, we identified several fake or incorrect references. However, the citation format gives the impression that the list is authoritative, particularly when a real journal or press is cited. This exercise of examining citation lists helps students to learn the value of checking sources, an essential skill when viewing LLMs as Frankfurtian bullshit machines.

Clarifying issues, conclusions, or beliefs. A key element of critical thinking is the identification and evaluation of facts and evidence supporting a claim. An LLM could be prompted to generate a list of facts needed to support a specific argument or viewpoint. Rather than debating the merits of the identified rationale, the students could be instructed to link them as provided into a proof of the claim; the discussion could then focus on whether the argument is logically consistent and what other pieces of information are still required.

Making interdisciplinary connections. We used this approach when asking about Dewey’s influence on Bernstein. Students could be given ostensibly unrelated topics from disparate fields and instructed to form queries to connect them in some way. The ensuing discussion could focus on whether the LLM identified valid connections; in particular, students could be asked to evaluate the output as a stochastic parrot by comparing

similarities in the input and output text. Students could also use the output as an initial prompt to find additional links that are less reliant on linking keywords.

Exercising fairmindedness. When discussing a controversial topic, students may come into a discussion with a strong pre-formed opinion. Rather than advocating for this *a priori* belief, the instructor could prompt the students to use an LLM to construct an argument of the opposite view. During a subsequent discussion, the students would be expected to use this output for the basis of their argument, relying on their peers to identify the argument's flaws.

Transferring insights to new contexts. After learning about a new skill, historical insight, or scientific property, an LLM could be asked to apply the concept to a different situation. Students could then be tasked with evaluating whether the transfer is valid, as well as identifying the flaws in the analogy. Our example of prompting the LLMs to generate an image of Dewey and Bernstein could be adapted for this purpose, as students could analyze both the image and the textual description to determine if the claimed similarities were valid.

These exercises illustrate several ways to link AI with the teaching of critical thinking. In some cases, we are directly asking the LLM for an output that will be discussed and evaluated. In others, we are using the LLM output to support other traditional approaches to critical thinking, such as constructing arguments or inspiring novel connections. The crux of all of these approaches, however, is that the instructor plays a central role in modeling proper use of the tool: the LLM output should always be used but should always be challenged. In particular, in all cases, the instructor could augment the discussion by asking one additional question: what accountability should be applied if the LLM results were used as provided.

7 Conclusions

AI tools—and LLMs, in particular—are popular tools and will continue to gain widespread use in education. Learning to challenge AI systems has become an essential part of the critical thinking learning outcomes that lie at the heart of education. Students will need to learn to approach LLM results critically and should not treat the output as an authoritative source. Our goal in this article has been to provide the necessary context and potential directions for using these tools successfully to augment the social process of education.

References

1. Agarwal, N., Moehring, A., Rajpurkar, P., and Salz, T.: Combining human expertise with artificial intelligence: experimental evidence from radiology. Nat. Bureau Econ. Res. Working Pap. Ser. <http://www.nber.org/papers/w31422>
2. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? From FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021)

3. Bergstrom, C.T., West, J.D.: Modern-day oracles or bullshit machines [online course] (n.d.). <https://thebullshitmachines.com/>.
4. Brookfield, S.D.: Teaching for Critical Thinking. Jossey-Bass (2012)
5. Byju, A.: The “Godfather of AI” predicted I wouldn’t have a job. He was wrong. The New Republic (2024, October 25). <https://newrepublic.com/article/187203/ai-radiology-geoffrey-hinton-nobel-prediction>.
6. Chen., B., Zhu, X., Díaz del Castillo, F.: Integrating generative AI in knowledge building. Comput. Educ. Artif. Intell. (2023). <https://doi.org/10.1016/j.caeai.2023.100184>
7. Chu, L.C., Anandkumar, A., Shin, H.C., Fishman, E.K.: The potential dangers of artificial intelligence for radiology and radiologists. J. Am. Coll. Radiol. JACR **17**(10), 1309–1311 (2020). <https://doi.org/10.1016/j.jacr.2020.04.010>
8. De Angelis, L., et al.: ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front. Public Health (2023). <https://doi.org/10.3389/fpubh.2023.1166120>
9. Frankfurt, H.: On Bullshit. Princeton University Press (2005)
10. Freeman, S., et al.: Active learning increases student performance in science, engineering, and mathematics. Proc. Natl. Acad. Sci. (PNAS) **111**(23), 8410–8415 (2014). <https://doi.org/10.1073/pnas.1319030111>
11. Gendron, Y., Andrew, J., Cooper, C.: The perils of artificial intelligence in academic publishing. Crit. Perspect. Account. (2022). <https://doi.org/10.1016/j.cpa.2021.102411>
12. Hicks, M. T., Humphries, J., and Slater, J.: ChatGPT is bullshit. Ethics Inf. Technol. (2024). <https://doi.org/10.1007/s10676-024-09775-5>
13. Hinton, G.: Comments at 2016 machine learning and market for intelligence conference (2016). <https://www.youtube.com/watch?v=2HMPRXstSvQ>
14. Hughes, A.: ChatGPT: everything you need to know about OpenAI’s GPT-4 tool (2023). www.sciencefocus.com. <https://www.sciencefocus.com/future-technology/gpt-3/>
15. Johnson, D.G.: Computer systems: moral entities but not moral agents. Ethics Inf. Technol. **8**, 195–204 (2006). <https://doi.org/10.1007/s10676-006-9111-5>
16. King, A.: From sage on the stage to guide on the side. Coll. Teach. (1993, Winter). <http://www.jstor.org/stable/27558571>
17. Kotek, H., Dockum, R., Sun, D.: Gender bias and stereotypes in large language models. Collective Intell. Conf. (CI ‘23) (2023). <https://doi.org/10.1145/3582269.3615599>
18. Kolata, G.: Doctors are using ChatGPT to improve how they talk to patients. The New York Times (13 June 2023). <https://www.nytimes.com/2023/06/12/health/doctors-chatgpt-artificial-intelligence.html>.
19. Lang, J.M.: Cheating Lessons: Learning from Academic Dishonesty. Harvard University Press (2013)
20. MacMillan, D., Ovalle, D., and Schaffer, A.: Arrested by AI: police ignore standards after facial recognition matches. Washington Post (13 Jan 2025). <https://www.washingtonpost.com/business/interactive/2025/police-artificial-intelligence-facial-recognition/>.
21. Murawski, L.M.: Critical thinking in the classroom...and beyond. J. Learn. High. Educ. (2014)
22. Nissenbaum, H.: Accountability in a computerized society. Sci. Eng. Ethics **2**, 25–42 (1996)
23. Oakley, B., Felder, R.M., Brent, R., Elhadj, I.: Turning student groups into effective teams. J. Student Centered Learn. **2**(1), 9–34 (2004)
24. Paul, R., Binker, A.J.A.: Critical thinking: what every person needs to survive in a rapidly changing world. Rev. 3rd ed. Foundation for Critical Thinking (1993)
25. Pettit, P.: Five elements of group agency. Inquiry **68**(3), 1069–1089 (2023). <https://doi.org/10.1080/0020174X.2023.2208391>
26. Piatek, S.J.: ChatGPT empowering non-English speakers (3 Jan 2023). www.sjpiatek.com. <https://www.sjpiatek.com/notes/chat-gpt-empowering-non-english-speakers/>

27. Sajid, H.: What are LLM hallucinations? Causes, ethical concern, & prevention. Unite.AI (09 April 2023). <https://www.unite.ai/what-are-llm-hallucinations-causes-ethical-concern-prevention/>.
28. Sakai, N.: Native, non-native, or bilingual? A concise assessment of ChatGPT's suitability for second-language instruction as a native or non-native pedagogue. OSF Preprints (31 March 2023). <https://doi.org/10.31219/osf.io/hy9ju>
29. Searle, J.: Minds, brains and programs. *Behav. Brain Sci.* **3**(3), 417–457 (1980). <https://doi.org/10.1017/S0140525X00005756>
30. Smith, C.: Hallucinations could blunt ChatGPT's success. *IEEE Spectrum* (13 March 2023). <https://spectrum.ieee.org/ai-hallucination>
31. Weimer, M.: *Learner-centered teaching: five key changes to practice*. Jossey-Bass (2013)
32. Yang, Y., Zhang, H., Gichoya, J.W., Katabi, D., Ghassemi, M.: The limits of fair medical imaging AI in real-world generalization. *Nat. Med.* **30**, 2838–2848 (2024). <https://doi.org/10.1038/s41591-024-03113-4>



AI Ethics in Higher Education Content Creation

Álvaro Carrasco-Aguilar¹(✉) , Miguel Camacho-Ruiz² ,
María Concepción Parra Meroño¹ , and María Mercedes Carmona Martínez¹

¹ Católica San Antonio de Murcia University of Murcia: Murcia, Murcia, Spain
alvarocarrascogr@gmail.com

² Complutense University of Madrid, Madrid, Spain

Abstract. Artificial Intelligence (AI) is rapidly transforming higher education, particularly in content creation, offering tools for personalized learning and automated resource generation. While this presents significant opportunities, it also raises critical ethical concerns regarding academic integrity, intellectual property, data privacy, bias, and the professor-student relationship. Addressing these challenges requires a robust framework to validate AI-generated educational content. This paper proposes a Delphi method-based methodology to achieve expert consensus on the ethical validation of such content. The Delphi method, with its iterative, anonymous, and feedback-controlled structure, is well-suited for navigating complex, novel issues lacking definitive answers. This study outlines the application of the Delphi process, from problem definition and expert panel selection to the development of questionnaires centered on key ethical criteria identified from the literature. The primary objective is to establish a validated set of guidelines and practical indicators for educators and institutions to ensure the responsible and ethical adoption of AI in educational content creation. This research aims to fill a critical gap by providing a structured approach to integrating AI responsibly into academic practices, safeguarding core educational values.

Keywords: Artificial Intelligence · Higher Education · Ethics

1 Introduction

Artificial Intelligence (AI) is rapidly transforming numerous sectors, and higher education is no exception to this technological revolution. Its increasing integration reshapes pedagogical dynamics, presenting a spectrum of possibilities alongside significant challenges. AI's capacity to process vast information and generate coherent responses makes it a potentially valuable tool for creating and adapting educational materials [1].

One of the areas of greatest AI impact in education is content generation. Tools based on generative AI, such as large language models (LLMs), allow for the rapid creation of texts, summaries, lesson plans, and even the formulation of assessment questions, offering educators novel methods to develop personalized and diverse learning resources [2]. This capability can reduce teacher workload, facilitate content adaptation to specific student needs, and promote more interactive and profound learning [1]. However, the

ease and speed of AI content generation raise important considerations. Adopting these technologies prompts fundamental questions regarding the quality, reliability, originality, and, crucially, the ethical suitability of the generated material [2]. Central concerns, including academic integrity, intellectual property rights, student data privacy, algorithmic bias perpetuation, and impacts on professor-student interaction, require proactive solutions [2].

Faced with the need to establish criteria and guidelines in a field as novel and multifaceted as the ethical validation of AI-generated content, where opinions may diverge and knowledge is still under construction, various research methodologies could be considered. For example, case studies could offer in-depth analyses of specific implementations, but their results would be difficult to generalize. Large-scale surveys could gather a wide range of opinions but would lack the deliberation and iterative refinement necessary to build consensus on complex issues. Focus groups or open discussion panels would allow for direct interaction and idea generation but run the risk of group dynamics being influenced by dominant personalities, pressure for immediate conformity, or the difficulty of reaching a structured consensus [3]. In this context, the Delphi method offers distinct advantages. Unlike the aforementioned methodologies, the Delphi method ensures participant anonymity, fostering honest and reflective expression of opinions, even if they are minority views, and mitigating biases from direct group interaction. Its iterative nature, with controlled feedback rounds, allows experts to reconsider and refine their judgments in light of the group's arguments and aggregate opinion, facilitating a gradual convergence towards a more robust and considered consensus than could be obtained through single surveys or less structured group discussions [4]. This ability to systematically aggregate and refine collective judgments makes it particularly suitable for exploring complex and ill-defined problems, where there is no single "correct" answer and the collective wisdom of a diverse panel of experts is sought for the formulation of criteria and guidelines.

Despite the potential benefits, the absence of a validated framework for the ethical and pedagogical evaluation of AI-generated educational content represents a critical gap. It is imperative to ensure that the material used in teaching-learning processes is not only accurate and relevant but also aligns with the fundamental ethical values and pedagogical principles that underpin quality education. In this context, the Delphi method, an expert consensus technique widely used in various disciplines, including education [5], emerges as an ideal methodological tool to address this emerging issue. Its ability to structure communication among experts and facilitate convergence towards informed agreements makes it a valuable instrument for developing guidelines and validation criteria in a field as novel and complex as AI in educational content creation.

The main objective of this article is to propose and detail a methodological framework, based on the Delphi method, for the ethical validation of educational content generated by Artificial Intelligence in the context of higher education. Specifically, this article aims to: 1) achieve expert consensus on key ethical criteria for evaluating AI-generated content, and 2) propose a structured procedure for institutions and educators to ensure responsible adoption of these technologies.

To achieve this objective, this article is structured as follows: Sect. 2, reviews the state of the art on the use of the Delphi method in education, highlighting its versatility and

robustness, and noting the existing gap in its application to the validation of AI-generated content. Section 3 describes the Delphi method in detail, its key characteristics, typical process phases, and some of its variations. Subsequently, Sect. 4 constitutes the core of our proposal, detailing how the Delphi method will be applied to validate AI-generated educational content, specifying the proposed study's phases, from problem definition and expert selection to questionnaire development and analysis to reach consensus on the identified ethical criteria. Finally, Sect. 5 will present the study's conclusions and outline future lines of research in this crucial area.

2 State of the Art: Use of the Delphi Method in Education

The Delphi method, originally conceived in the 1950s by Olaf Helmer and Norman Dalkey at the RAND Corporation, emerged as a systematic technique for prediction and obtaining consensus among experts, especially in contexts where information was uncertain or incomplete, such as military defense planning [6, 7]. It is founded on the premise that structured group decisions from experts are often more accurate and reliable than individual judgments, leveraging the principle that “two heads are better than one” [7]. Since its beginnings in technological and military forecasting, the applicability of the Delphi method has extended to a vast range of disciplines, including public administration, healthcare, business, and, significantly, education [8, 9].

In the educational field, the Delphi method has proven to be a versatile and valuable tool. As early as the 1970s, its first applications in higher education administration were documented for strategic planning, defining institutional objectives, curriculum design, and resource allocation [10]. Educational researchers and planners soon recognized its potential to structure communication among various stakeholders (teachers, administrators, students, employers) and facilitate informed decision-making in complex scenarios.

Subsequent studies solidified Delphi's application across multiple educational facets. For example, it has been employed to identify key competencies needed in vocational training programs, such as in pharmaceutical education, where it has helped define the knowledge, skills, and attitudes that future pharmacists should possess [11, 12]. It has also been fundamental in curriculum review and updating, allowing institutions to adapt their programs to the changing needs of the labor and social environment [7, 11]. Furthermore, the Delphi method has been used to establish institutional priorities, identify teacher training needs, and evaluate the relevance of new pedagogies or educational technologies [9, 13]. In the context of medical education, for example, Delphi has been used to develop consensus on digital health competencies that should be included in the training of future doctors [14]. More recently, its application has extended to exploring barriers and facilitators for the educational success of specific student groups, such as those learning English as an additional language [15].

The method's robustness in anonymously and iteratively aggregating expert judgments makes it ideal for controversial or multifaceted topics common in educational planning and policy, where single correct answers are rare. Despite the wide range of applications of the Delphi method in education, spanning from defining objectives to program evaluation and identifying future trends, a review of current literature has not

revealed studies specifically employing the Delphi method to address the emerging challenge of validating and ethically controlling educational content generated by Artificial Intelligence. This gap highlights a critical and novel opportunity for the Delphi method to provide a structured framework for achieving expert consensus on how to responsibly integrate these new technologies into educational content creation processes.

3 Key Characteristics of the Delphi Method

The Delphi method is fundamentally a structured communication process, designed to obtain, aggregate, and refine the informed judgment of a group of experts on a specific topic, aiming to reach a consensus or identify significant points of divergence [6, 13]. Its development was based on the idea that structured group forecasts or decisions are, in general, more reliable than unstructured individual opinions. Unlike traditional group discussions, Delphi seeks to minimize biases associated with direct interaction, such as the influence of dominant personalities, group pressure, or semantic “noise” [6, 7].

The Delphi methodology is distinguished by several essential characteristics that ensure the systematicity and quality of the consensus-building process [9, 13]:

1. **Anonymity of Participants:** Experts provide their judgments individually and anonymously. A facilitator or research team centralizes, processes, and disseminates anonymized responses to the panel, protecting contributor identities. This anonymity fosters honesty, reduces inhibition, and minimizes the influence of authority or status of certain participants over others [6, 9].
2. **Iteration:** The Delphi process unfolds through multiple rounds of consultation. In each round, experts are presented with a questionnaire (which may include open-ended questions in the first round and rating scales or closed-ended questions in subsequent rounds). After each round, responses are analyzed, and feedback is provided to the panel [6, 11].
3. **Controlled Feedback:** After each round (except the last), experts receive a summary of the previous round’s results. This feedback typically includes measures of central tendency (such as the median or mean of group responses), the degree of dispersion (such as the interquartile range), and, in some cases, the most relevant minority or majority arguments. This allows each expert to reconsider their judgment in light of the collective opinion [9, 13].
4. **Statistical Group Response:** The final output of the Delphi process is not merely an individual’s opinion but a statistical aggregation of the panel’s final judgments. The goal is for group responses to converge through iteration and feedback, although persistent disagreements can also be reported if total consensus is not achievable or desirable for the problem in question [6, 9].

4 Phases of the Delphi Process

Although variations exist, a typical Delphi study follows a sequence of well-defined phases [4, 6, 8]:

1. **Problem Definition and Study Design:** The problem to be investigated and the objectives of the Delphi study are clearly defined. The protocol is designed, including the research question, the definition of “expert” for the context, criteria for panel selection, the estimated number of rounds, and criteria for defining consensus.
2. **Expert Panel Selection:** A group of individuals with relevant knowledge and experience in the study topic is identified and recruited. Panel size may vary, but a balance between diversity of perspectives and process manageability is generally sought.
3. **First-Round Questionnaire Development:** If starting from scratch, the first round usually consists of open-ended questions for experts to generate ideas, identify key factors, or propose solutions. Alternatively, if a body of knowledge already exists, items based on literature can be presented for initial assessment.
4. **Conducting Iterative Rounds**
 - **Round 1:** The initial questionnaire is sent to panelists. Responses are collected and analyzed.
 - **Subsequent Rounds (2, 3,...n):** A new questionnaire is developed based on the results of the previous round. This questionnaire includes statistical feedback (medians, ranges, etc.) and, sometimes, summaries of arguments. Experts are asked to reconsider their previous responses. If their responses deviate significantly from the group trend, they may be asked to justify their stance. This process is repeated until a pre-established level of consensus is reached or stability in responses is observed.
5. **Consensus Analysis and Definition:** In each round, and especially in the final one, data are analyzed to determine the degree of agreement. Consensus does not always imply unanimity; it is often defined as a percentage of agreement (e.g., 70–80%) on a rating scale or a significant reduction in response variance.
6. **Final Report Elaboration:** The methodology, results of each round, level of consensus reached for each item, and study conclusions are documented.

4.1 Variations of the Delphi Method

The “classic” Delphi method has evolved, leading to various adaptations to suit different needs and contexts. The e-Delphi uses online platforms for questionnaire administration and feedback, streamlining the process and facilitating participation from geographically dispersed experts [11]. The Policy Delphi does not necessarily seek consensus but focuses on exploring and clarifying the full spectrum of opinions and opposing arguments on a complex policy issue, identifying points of agreement and disagreement [7].

A particularly relevant variant in contexts of uncertainty and subjectivity is the Fuzzy Delphi Method (FDM). The FDM integrates fuzzy set theory with the traditional Delphi procedure to handle the vagueness and imprecision inherent in human judgment, especially when using linguistic scales (e.g., “very important,” “not very important”) [8]. Instead of precise numerical responses, experts can express their opinions using fuzzy numbers (such as triangular or trapezoidal numbers) that represent a range of possible values with different degrees of membership. This approach enables a more realistic modeling of uncertainty and can yield more robust and nuanced consensus, proving especially useful for selecting criteria, evaluating alternatives, or forecasting in complex environments [8].

5 Delphi Method Applied to the Validation of AI-Generated Educational Content

The integration of AI in higher education has transformed content creation, enabling tools for personalized learning, adaptive assessments, and automated resource generation. In the evolving educational landscape, professors are increasingly embracing content creation to enhance teaching and learning experiences [16]. By developing customized digital materials, such as interactive modules, videos, and podcasts, educators can tailor content to meet the specific needs of.

their students, thereby fostering deeper engagement and understanding [17]. This advancement, however, raises critical ethical questions concerning academic integrity, intellectual property rights, data privacy and security, bias and fairness, and its impact on professor-student relationships [18]. These core ethical challenges were selected based on their consistent identification as critical, recurring concerns in contemporary literature and ethical guidelines on AI in higher education [18, 20, 22, 24, 25, 28]. Navigating this challenge requires a balanced approach that leverages AI's potential without compromising core academic values. This study proposes a strategy for responsible adoption, incorporating expert consensus methodologies such as the Delphi method [19].

As detailed in Sect. 2, the Delphi method offers a robust framework for anonymously and iteratively aggregating expert judgments, making it ideal for addressing topics for which there is no single correct answer. While broadly applied in education, no studies were found that specifically use the Delphi method for validating and ethically controlling AI-generated educational content. This section describes how the Delphi method will be applied to address this gap, following its traditional phases.

5.1 Phase 1: Problem Definition and Study Design

This initial phase is crucial for establishing the foundations of the Delphi study. It involves precisely defining the problem of ethical validation of AI-generated content, formulating the study objectives, and designing the general protocol, including the identification of key criteria that experts will evaluate.

The specific problem to be investigated is the need for a consensual framework for the ethical and pedagogical validation of educational content generated by AI before its implementation in higher education environments.

The main objectives of this Delphi study are:

1. To identify and achieve expert consensus on the fundamental ethical criteria for evaluating AI-generated educational content.
2. To validate the relevance and applicability of ethical criteria identified in the literature (academic integrity, intellectual property, privacy, bias, and relational impact) for the practical evaluation of such content.
3. To obtain consensual recommendations on best practices and guidelines to mitigate ethical risks associated with the use of AI in educational content creation.
4. To propose a set of specific indicators or questions that educators and institutions can use to evaluate AI-generated content from an ethical perspective.

The study design will apply the Delphi method's key characteristics (anonymity, iteration, controlled feedback, and statistical group response, detailed in Sect. 3) to evaluate AI-generated content against the following literature-identified ethical criteria, which will guide expert evaluation:

- **Criterion 1:** Academic Integrity and Authenticity. AI's capacity to generate essays, lesson plans, and research summaries challenges traditional notions of authorship and originality. While tools like ChatGPT can enhance efficiency, overreliance risks diminishing the educator's critical role in crafting pedagogically rigorous materials [20]. Studies emphasize the need for clear guidelines to distinguish AI-assisted content from purely human-generated work, ensuring transparency and preserving academic rigor [21]. For example, universities are urged to update honor codes to address AI-generated plagiarism and define attribution standards for hybrid outputs, while policymakers are encouraged to adopt adaptive regulatory frameworks, such as pilot testing for AI governance, to balance innovation with accountability [22].
- **Criterion 2:** Intellectual Property Rights. AI systems trained on vast datasets, often including copyrighted materials, blur ownership boundaries. Current legal frameworks do not adequately resolve ownership of AI-generated content, whether it belongs to developers, users, or institutions [23]. Ethical governance of AI requires collaboration between academic institutions and regulatory bodies. Beyond institutional policies, mechanisms like regulatory sandboxes and policy prototypes enable safe testing of AI frameworks before formal adoption, ensuring alignment with evolving technological and ethical standards. Such approaches allow universities to trial attribution models and plagiarism detection tools in controlled environments, reducing risks to academic integrity [22].
- **Criterion 3:** Data Privacy and Security. Generative AI has the potential to enhance education, but it also raises ethical concerns related to bias, misinformation, and data privacy. Without adequate regulations and oversight, its use may exacerbate inequalities and lead to the misuse of sensitive information [24]. Ethical AI use mandates compliance with regulations like GDPR and FERPA, alongside transparent informed consent protocols. Institutions should implement technical measures (encryption, anonymization) and third-party audits to mitigate risks [25].
- **Criterion 4:** Bias and Fairness. AI algorithms often perpetuate biases embedded in training data, leading to culturally insensitive or exclusionary content. For instance, generative models may underrepresent non-Western perspectives in educational materials [26]. To promote equity, regular bias audits, diverse dataset curation, and inclusive design practices are essential. Educators must complement AI tools with critical pedagogy to counteract algorithmic limitations [27].
- **Criterion 5:** Impact on Professor-Student Relationships. While AI personalizes learning, excessive automation risks reducing direct interaction between professors and students, eroding trust and hindering critical thinking development [28]. Hybrid models that combine AI efficiency with human oversight, such as AI-generated drafts refined by educators, can preserve essential relational dynamics [29].

These five criteria will form the basis for developing the first-round questionnaire.

5.2 Phase 2: Expert Panel Selection

A multidisciplinary expert panel will be identified and recruited. Selection criteria will include demonstrated experience in areas such as pedagogy in higher education, development or use of AI in educational contexts, AI ethics, and digital law applied to education. A diverse representation of profiles (academics, technologists, ethics specialists, jurists) will be sought to enrich the evaluation from multiple perspectives. Panel size will be determined to ensure diverse opinions while maintaining process manageability.

5.3 Phase 3: First-Round Questionnaire Development

An initial questionnaire will be designed to present experts with an example (or several examples) of educational content generated by a specific AI tool (e.g., a learning module, a set of assessment questions, a thematic summary). The questionnaire will be structured around the five key ethical criteria defined above. For each criterion, questions will be formulated as follows:

- Closed-ended questions (using Likert scales, e.g., from 1 to 5 or 1 to 7) for experts to assess the extent to which AI content meets or presents risks regarding that criterion.
- Open-ended questions for experts to justify their assessments, identify nuances, point out specific uncovered concerns, and propose improvements or guidelines.

For example, under the criterion “Academic Integrity and Authenticity,” one might ask: “To what extent do you consider this AI-generated content to respect principles of academic originality?” (Likert scale), followed by: “Please justify your answer and point out any specific aspect of the content that concerns you in terms of academic integrity.”

5.4 Phase 4: Conducting Iterative Rounds

- **Round 1:** The initial questionnaire will be distributed to the expert panel. Upon receipt of responses, they will be analyzed quantitatively (descriptive statistics for Likert scales) and qualitatively (content analysis for open-ended responses).
- **Subsequent Rounds** (2–3 rounds are anticipated in total): A second questionnaire will be developed. This will include feedback from Round 1:
- The medians (or means) and interquartile range of group ratings for each Likert scale item.
- An anonymized summary of the main arguments, justifications, and concerns expressed in open-ended responses. Experts will be asked to reconsider their initial ratings in light of the group’s response and the arguments presented. Experts whose revised opinions still diverge significantly from the group trend may be asked for further justification. This iterative process will continue until a predefined level of consensus is reached or stability in responses is observed.

5.5 Phase 5: Consensus Analysis and Definition

Consensus will be defined a priori, for example, as a minimum percentage of agreement (e.g., $\geq 75\%$ of experts scoring in the top two values of a 5-point Likert scale for a

specific item) or by the stability of medians and a reduction in dispersion (interquartile range) of responses between rounds. Both points of consensus and areas of persistent disagreement will be analyzed, documenting key arguments for both.

5.6 Final Phase 6: Final Report Elaboration

AI's introduction into higher education, particularly for content generation, offers promising innovation but also demands significant ethical reflection and robust validation mechanisms. This article addresses the challenge of ensuring that AI-generated educational content aligns with ethical principles, proposing a Delphi-based framework for validation. A literature review identified five key ethical criteria for evaluating such content: academic integrity and authenticity, intellectual property rights, data privacy and security, bias and fairness, and the impact on professor-student relationships.

6 Conclusions and Future Work

This work's primary contribution is a detailed proposal for applying the Delphi method, a collaborative, systematic approach, to achieve expert consensus on the ethical suitability of AI-generated content. The Delphi method's characteristics, including anonymity, iterative feedback, and statistical aggregation, make it an ideal tool for tackling the complex challenges posed by AI. The framework not only validates the identified ethical criteria but also offers practical guidelines and specific indicators to help educators and institutions.

We conclude that ethically validating AI-generated educational content is essential for ensuring quality, equitable education that preserves core academic values. Adopting AI without critical reflection could lead to significant risks, such as biased or inaccurate information, harm to academic integrity, and the dehumanization of learning. The Delphi method offers a structured approach to mitigate these risks, thereby fostering responsible AI integration in educational content creation.

This methodological proposal paves the way for conducting the Delphi study, with the next steps focusing on recruiting an expert panel, designing the first-round questionnaire, and executing iterative rounds until consensus is reached. The study's results will provide expert-validated guidelines, contributing to the educational community.

Future research should focus on developing and validating specific rubrics, conducting longitudinal studies on how perceptions and ethical criteria evolve with AI technology, and researching the practical implementation of the guidelines across various educational contexts. Additionally, AI training and literacy programs for educators should be developed, and student perspectives should be incorporated into future studies on AI-generated content. Expanding research to diverse educational contexts and cultural settings will also be crucial for adapting and validating ethical criteria inclusively.

Acknowledgments. This research was funded by the Complutense University Research Project titled "Impacts of Artificial Intelligence on the Tourism Industry from a Social Science Perspective" (project number 506-2024), as well as the Telefónica Chair on Smart Cities at Universitat Rovira i Virgili and Universitat de Barcelona (project number 42.DB.00.18.00).

Disclosure of Interests. The authors declare no competing interests.

References

1. Pal, S., Kumari, K., Kadam, S., Saha, A.: The AI Revolution. IARA Publication (2023)
2. Attard, A., Dingli, A.: Empowering educators: Leveraging large language models to streamline content creation in education. In: ICERI2024 Proceedings, pp. 1312–1321. IATED (2024)
3. Stewart, D.W., Shamdasani, P.N.: Focus groups: Theory and practice. Sage publications (2014)
4. Humphrey-Murto, S., et al.: The use of the Delphi and other consensus group methods in medical education research: a review. *Acad. Med.* **92**(10), 1491–1498 (2017)
5. Zawacki-Richter, O.: Research areas in distance education: a Delphi study. *Int. Rev. Res. Open Distribut. Learn.* **10**(3) (2009)
6. Dalkey, N.C.: The Delphi Method: An Experimental Study of Group Opinion. RAND Corporation, Santa Monica, CA (1969)
7. Boberg, A.L., Morris-Khoo, S.A.: The Delphi method: a review of methodology and an application in the evaluation of a higher education program. *Canadian J. Program Evaluat.* **7**(1), 27–39 (1992)
8. Ciptono, A., Setiyono, S., Nurhidayati, F., Vikaliana, R.: Fuzzy Delphi method in education: a mapping. *J. Phys.: Conf. Ser.* **1360**(1), 012029 (2019)
9. Humphrey-Murto, S., Wood, T.J., Gonsalves, C., Mascioli, K., Varpio, L.: The delphi method. *Acad. Med.* **95**(1), 168 (2020)
10. Judd, R.C.: Use of Delphi methods in higher education. *Technol. Forecast. Soc. Chang.* **4**(2), 173–186 (1972)
11. Olsen, A.A., Wolcott, M.D., Haines, S.T., Janke, K.K., McLaughlin, J.E.: How to use the Delphi method to aid in decision making and build consensus in pharmacy education. *Curr. Pharm. Teach. Learn.* **13**(10), 1376–1385 (2021)
12. Yeh, J.S., Van Hoof, T.J., Fischer, M.A.: Key features of academic detailing: development of an expert consensus using the Delphi method. *Am. Health Drug Benefits* **9**(1), 42 (2016)
13. Linstone, H.A., Turoff, M. (eds.): The Delphi Method: Techniques and Applications. Addison-Wesley Publishing Company, Reading, MA (1975)
14. Khurana, M.P., Raaschou-Pedersen, D.E., Kurtzhals, J., Bardram, J.E., Ostrowski, S.R., Bundgaard, J.S.: Digital health competencies in medical school education: a scoping review and Delphi method study. *BMC Med. Educ.* **22**(1), 129 (2022)
15. Oxley, E., Nash, H.M., Weighall, A.R.: Consensus building using the Delphi method in educational research: a case study with educational professionals. *Int. J. Res. Method Educ.* **48**(1), 29–43 (2025)
16. Bakar, S.: Investigating the dynamics of contemporary pedagogical approaches in higher education through innovations, challenges, and paradigm shifts. *Soc. Sci. Chronicle* **1**(1), 01–19 (2021). <https://doi.org/10.56106/ssc.2021.009>
17. Thelma, C.C., Sain, Z.H., Mpolomoka, D.L., Akpan, W.M., Davy, M.: Curriculum design for the digital age: Strategies for effective technology integration in higher education. *Int. J. Res.* **11**(07), 185–201 (2024). <https://doi.org/10.5281/zenodo.13123899>
18. Saz-Pérez, F., Pizà-Mir, B.: Needs and perspectives on the integration of generative artificial intelligence in Spanish educational context. *Necesidades y perspectivas de la integración de la inteligencia artificial generativa en el contexto educativo. Revista UTE Teaching & Technology (Universitas Tarraconensis)* (2), e3803 (2024)
19. López-Gómez, E.: El método Delphi en la investigación actual en educación: una revisión teórica y metodológica. *Educación Xx1*, 21(1), 17–40 (2018)

20. Bin-Nashwan, S.A., Sadallah, M., Bouteraa, M.: Use of ChatGPT in academia: academic integrity hangs in the balance. *Technol. Soc.* **75**, 102370 (2023)
21. Porsdam Mann, S., et al.: (2024). Guidelines for ethical use and acknowledgement of large language models in academic writing. *Nat. Machine Intell.* 1–3
22. UNESCO: Recommendation on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization (2021)
23. Delgado, N., Carrasco, L.C., de la Maza, M.S., Etxabe-Urbieto, J.M.: Aplicación de la Inteligencia Artificial (IA) en Educación: los beneficios y limitaciones de la IA percibidos por el profesorado de educación primaria, educación secundaria y educación superior. *Revista electrónica interuniversitaria de formación del profesorado* **27**(1), 207–224 (2024)
24. Wu, C., Zhang, H., Carroll, J.M.: AI Governance in Higher Education: Case Studies of Guidance at Big Ten Universities (2024). arXiv preprint [arXiv:2409.02017](https://arxiv.org/abs/2409.02017)
25. Oye, E., Frank, E., Owen, J.: Ethical Considerations in AI-Driven Education (2024)
26. Nyaaba, M., Wright, A., Choi, G.L.: Generative AI and digital neocolonialism in global education: Towards an equitable framework (2024). arXiv preprint [arXiv:2406.02966](https://arxiv.org/abs/2406.02966)
27. Balan, A.: AI and Legal Education: Ethical and Sustainable Approaches. Taylor & Francis (2025)
28. George, A.S.: Technology tension in schools: addressing the complex impacts of digital advances on teaching, learning, and wellbeing. *Partners Univ. Multidisc. Res. J.* **1**(3), 49–65 (2024)
29. Paciello, L.: Human capital efficiency: a comprehensive overview (2024)



GenAI and Proportionality: European and Portuguese Ethical-Legal Framework

Filipa Pais D'Aguiar¹ , Fernanda Duarte² , and Nuno Silva³  

¹ CEJEIA, Lusíada University, Lisbon, Portugal
13001224@lis.ulusiada.pt

² INCT-InEAC, Federal Fluminense University, Niterói, Brazil
fernanda_duarte@id.uff.br

³ COMEGI, Lusíada University, Lisbon, Portugal
nsas@ulusiada.pt

Abstract. Generative AI systems are changing society, the way we relate with each other and the way we perceive time, space, and reality. The research problem we now focus on addresses the still unsettled and yet-to-be-found Human-AI relation keystone. Firstly, how AI systems regulation complies with explainability, transparency and accountability demand as well as with a responsible, educated, conscious, ethical and trustworthy usage. Secondly, how the definition of “high-risk” AI systems may impact Human-AI sustainable development. Thirdly, ponder the proportionality principle framework, within the scope of the EU AI Act and the GDPR, considering special categories of personal data. Although the focus of this research is on Portuguese and European Union ethical-legal framework solutions on all these issues, we have adopted comparative law methodology with the dialogue and contributions from other ethical-legal orders.

Keywords: GenAI · Human Rights · Proportionality

1 Introduction

Artificial Intelligence systems (henceforth, AI) and, particularly, Generative AI systems (henceforth, GenAI), are changing society, the way we relate with each other and the way we perceive time, space, and reality. The main challenges and benefits that arise from AI systems usage have been the aim of previous studies and led us to the current research proposal.

The research problem we now focus on addresses the still unsettled and yet-to-be-found Human-AI relation keystone, meaning, looking forward to measures that may contribute to strengthen the thin balance between regulation, compliance and Human-AI sustainable development [1]. Although this research focus on Portuguese and European Union ethical-legal framework solutions on these issues [2], the dialogue and contributions from other ethical-legal orders (*e.g.* Brazil and Chile) will be referred to in an innovative open-minded and comparative law perspective. Hence, we have adopted the comparative law methodology, particularly, the micro-comparison methodology,

the hypothetical-deductive methodological approach and historical methodological procedure on global inequalities and digital justice.

Therefore, the development of the research problem is structured around three main hypotheses: first, examining how the regulation of AI systems may comply with Human Rights standards; second, reflecting on the definition of 'high-risk' AI systems and its implications for sustainable development; and third, analysing the praxis of the proportionality principle in the regulation of personal data, particularly within the frameworks of the EU AI Act and the EU General Data Protection Regulation (henceforth, GDPR). Lastly, the study aims to explore proposals that foster a balanced approach between Human Rights protection, the proportionality principle, and AI systems regulation.

2 Human Rights and AI Systems

Firstly, in examining how the regulation of AI systems may comply with Human Rights standards, it is understood that such regulation must be aligned with core democratic and legal principles—namely, explainability, transparency, and accountability—as well as with a broader normative commitment to the responsible, informed, conscious, ethical, and trustworthy use and dissemination of these technologies. The integration of these principles into AI governance is not merely aspirational; it is a legal and ethical necessity in democratic societies that seek to preserve individual dignity, equality, and freedom in the face of rapidly evolving technological capabilities. Effective regulation must therefore move beyond technical oversight to encompass normative safeguards that ensure algorithmic decision-making does not reinforce discrimination, infringe upon privacy, or erode procedural fairness. In this sense, the design and deployment of AI systems should be guided by a rights-based approach, one that embeds safeguards throughout the AI lifecycle—from data collection and model training to praxis and monitoring. This requires not only robust legal frameworks, but also multidisciplinary collaboration among policymakers, technologists, ethicists, and civil society to anticipate harms, enhance public trust, and ensure that innovation remains compatible with fundamental rights.

As artificial intelligence continues to permeate high-stakes domains—ranging from public administration and healthcare to criminal justice and education—the legal and ethical implications of autonomous decision-making systems are becoming increasingly intricate and contested [3]. The problem of accountability in AI-driven actions, particularly in the context of generative AI (GenAI)¹, highlights urgent concerns related to algorithmic opacity, embedded biases, and the risk of unintended or harmful consequences [4, 5]. These challenges are not only technical but fundamentally normative: opaque systems compromise due process and democratic oversight, especially when deployed in contexts where individuals' rights and freedoms are at stake.

¹ Generative AI (GenAI) refers to a class of artificial intelligence systems capable of generating new content—such as text, images, audio, video, or synthetic data—based on patterns learned from training datasets. These systems do not merely analyse or classify existing data; they create original outputs that resemble human-generated content. For further information, see the OECD Framework for the Classification of AI Systems.

As demonstrated in the evolving legislative debates around the European Union's AI Act [6, 7], the regulatory challenge is twofold: it requires both a precise categorization of high-risk AI uses and the implementation of enforceable mechanisms for ensuring algorithmic transparency and explainability. As [8] notes, algorithmic decision-making introduces a new layer of epistemic and political complexity that demands a rethinking of conventional regulatory paradigms.

In this context, compliance with human rights standards cannot be viewed as a mere add-on to innovation policy; it must be structurally integrated into the design, deployment, and monitoring of AI systems. This demands a robust, multidisciplinary governance framework in which legal experts, ethicists, data scientists, engineers, and public stakeholders collaboratively articulate the boundaries of legitimate and illegitimate uses of AI [9]. Such a framework must also ensure that regulatory measures do not stifle technological innovation but rather anchor it in a socially sustainable and normatively sound foundation. Legal regulation, therefore, should not be conceptualized as a barrier to development but as a cornerstone – a structural enabler of both AI-driven innovation and the safeguarding of fundamental rights. This balance is critical to maintaining public trust and legitimacy in the governance of artificial intelligence.

3 “High-risk” AI Systems and Sustainable Development: Neuro-Rights

Secondly, this research explores how the definition of ‘high-risk’ AI systems may shape the trajectory of sustainable development, particularly in terms of ensuring that human rights remain central to AI design and deployment. The challenges faced during the preparatory legal work and the technical meetings, concerning the EU AI Act development and its definition of “high-risk” systems, particularly through 2023 and 2024, clearly mirrored the difficulties arising from the thin balance between Human Rights protection, AI regulation, the proportionality principle and the demanding velocity of AI innovation, digital skills and competences [10].

For instance, on the one hand, a wide set of exceptions for biometric data usage by law enforcement was considered, within the scope of EU AI Act, on the argument that law enforcement should not be prevented to fully perform their activity by the EU AI Act regulation [6]. On the other hand, the protection of neuro-rights, privacy and psychological integrity must also be considered as we are witnessing the very real possibilities of hacking the human brain and the human thought [11–13], for instance, through non-authorized usage of data collected from neuroimaging techniques such as computed tomography, positron emission tomography (PET), single-photon emission computed tomography (SPECT), functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) [14, 15].

Therefore, there are potential risks of mental hacking resulting not only from the usage of data collected by neuroimage techniques but also from the interaction between Humans and AI and how it may influence Human decision-making processes (*e.g.*, the usage of AI in courts, by judges, in a support level) [16, 17].

But one thing is the ability to copy or clone an original human thought that is unique to that specific human being and doing it by using the information collect from

neuroimaging techniques. Other thing (more difficult to identify) is the ability to modify, *ab initio*, the human thought, i.e., suggestibility. Suggestibility is far from being unknown [18]². However, the source is now rather different. AI systems may now be a source for modifying human memories. For [19], the era of fake news does rely in the lack of rigor in the register of human memory as human neural systems that support memory are not static but malleable. And one must consider that human freedom (and, more importantly, freedom of choice) is determined by the laws of nature that act in our brain [20, p. 67]; also, [11, p. 113]. So, to protect this specific vulnerability of human thought and to prevent mental hacking (either by copying original human thought or by influencing it), several legislative initiatives were put in force.

For instance, back in 2021, the Constitution of Chile was modified to specifically protect the brain activity and the information resulting from it [15]³. In Brazil, the Bill No. 522/22⁴, emphasize, in 2022, the relevance of protecting neuro-rights, while in 2023 the legislative process for the development of a Brazilian Artificial Intelligence Act remains underway [21, 22]⁵. In the European Union, the EU AI Act art. ° 5, n. ° 1, par. a) prohibits

² The American Psychology Association defines suggestibility by “1. an inclination to readily and uncritically adopt the ideas, beliefs, attitudes, or actions of others. 2. an occasional synonym for hypnotic susceptibility. [...]”.

³ Constitution of Chile, 25 October 2021, Law n.° 21.383, that modifies former n. ° 1, art. ° 19.° and establishes that: “[...]”El desarrollo científico y tecnológico estará al servicio de las personas y se llevará a cabo con respeto a la vida y a la integridad física y psíquica. La ley regulará los requisitos, condiciones y restricciones para su utilización en las personas, debiendo resguardar especialmente la actividad cerebral, así como la información proveniente de ella; [...]”.

⁴ Legislative proposals on neurorights in Brazil aim to safeguard fundamental rights related to the human brain and nervous system, ensuring autonomy, privacy, and mental integrity in the face of emerging neurotechnologies. Bill (projeto de lei) No. 2174/2023 introduces the concept of “neurorights” as a set of fundamental protections against the misuse of neural data and its potential medical, ethical, and social consequences. Bill (projeto de lei) No. 522/2022 proposes amendments to Brazil’s General Data Protection Law (LGPD) to explicitly include neural data—defined as information obtained directly from the brain—under its scope. Together, these initiatives seek to: (i) uphold individual rights concerning the brain and nervous system; (ii) prevent unauthorized manipulation of neural activity and protect free will; (iii) avoid discrimination and reinforce mental privacy; and (iv) ensure equitable governance of neuroscience-driven technologies. The debate around neurorights represents in Brazil a novel legal frontier that addresses the ethical and normative challenges posed by the intersection of neuroscience, data protection, and human dignity in an increasingly technologized world.

⁵ Brazil’s Artificial Intelligence legislation is currently under development and public debate. In December 2024, the Federal Senate approved Bill (projeto de lei) No. 2.338/2023, which seeks to establish a regulatory framework for AI based on principles of safety, transparency, and accountability. The bill, currently under consideration in the Chamber of Deputies, lays out foundational provisions regarding AI regulation, including guiding principles, individual rights, and institutional responsibilities. It also proposes the creation of a National System for AI Regulation and Governance, aligning domestic oversight with international best practices. Notably, the bill categorizes certain AI systems as “high-risk” and subjects them to stricter requirements, particularly in sensitive domains such as public security, healthcare, and the judiciary. It further enshrines individual rights related to data protection, privacy, and access to

AI practices susceptible to manipulate human thought⁶. Also, the definition of “high-risk” AI systems provides a classification of AI systems that both protect human rights and look forward to finding a balance between Human-AI sustainable development.

The “high-risk” AI systems definition in the EU AI Act follows the previous version proposed, back in 2021, by the European Commission in the art. ° 6.° and Annex III [6]⁷. Also, the OECD, in 2022, provides a framework for the classification of AI systems [23] and the scholarship also seems to follow. For instance, in 2023, [24, p. 229] mentions a four based criterion for AI systems classification based on the low, medium, high and unacceptable risk they may represent for Human Rights. In the case of unacceptable risk, the system should be banned.

As for the EU AI Act consolidated version, the high-risk AI system definition can be apprehended under the art. ° 6.° combined with the dispositions of the Annex III. The art. ° 6.°, n. ° 1 describes the classification rules for high-risk AI systems and the art. ° 6.°, n. ° 2⁸ clarifies that “In addition to the high-risk AI systems referred to in paragraph 1, AI systems referred to in Annex III shall be considered to be high-risk.” [6, p. 53]. However, considering the time needed for the General-Purpose AI (henceforth, GPAI)

information. Transparency and safety in AI system design and deployment are core tenets of the proposal, which also delineates civil liability rules applicable to developers and users in the event of harm caused by AI systems.

⁶ EU AI Act (European Parliament and Council, 2024, p. 51): “Article 5 – Prohibited AI practices. 1. The following AI practices shall be prohibited: (a) the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person’s consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm; [...]”.

⁷ European Commission, 2021, art. ° 6.°. “1. Irrespective of whether an AI system is placed on the market or put into service independently from the products referred to in points (a) and (b), that AI system shall be considered high-risk where both of the following conditions are fulfilled: (a) the AI system is intended to be used as a safety component of a product, or is itself a product, covered by the Union harmonisation legislation listed in Annex II; (b) the product whose safety component is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on the market or putting into service of that product pursuant to the Union harmonisation legislation listed in Annex II. 2. In addition to the high-risk AI systems referred to in paragraph 1, AI systems referred to in Annex III shall also be considered high-risk.”.

⁸ EU AI Act (European Parliament and Council, 2024, p. 53). “Article 6 – Classification rules for high-risk AI systems. 1. Irrespective of whether an AI system is placed on the market or put into service independently of the products referred to in points (a) and (b), that AI system shall be considered to be high-risk where both of the following conditions are fulfilled: (a) the AI system is intended to be used as a safety component of a product, or the AI system is itself a product, covered by the Union harmonisation legislation listed in Annex I; (b) the product whose safety component pursuant to point (a) is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment, with a view to the placing on the market or the putting into service of that product pursuant to the Union harmonisation legislation listed in Annex I. [...]”.

providers to adopt all the measures, the art. ° 6.°, n. ° 1 will only enter into force from 2 August 2027, according to art. ° 113.° C, EU AI Act.

This delay in entering into force is justified by the need for the “Providers of general-purpose AI models that have been placed on the market before 2 August 2025 shall take the necessary steps in order to comply with the obligations laid down in this Regulation by 2 August 2027”, according to art. ° 111.°, n. ° 3. Also, the “AI systems which are components of the large-scale IT systems established by the legal acts listed in Annex X that have been placed on the market or put into service before 2 August 2027 shall be brought into compliance with this Regulation by 31 December 2030.”, according to art. ° 111. °, n. ° 1, EU AI Act.

So, the question is how to act meanwhile? Maybe, and apart from relevant legislation already in force (*e.g.*, GDPR; Chapters I and II of the EU AI Act that entered into force from the 2 February 2025; art. ° 56.°, n. ° 9, on the Codes of Practice for GPAI, that entered into force from the 2 May 2025), the answer may also be found on the protection provided by the proportionality principle and the scholarship that follows (*e.g.*, [25]).

4 Proportionality Principle Framework and Personal Data: EU AI Act and GDPR

Thirdly, we turn towards an analysis of the proportionality principle framework, within the scope of the EU AI Act and the GDPR, considering special categories of personal data. First, the question raised is: how does the proportionality principle answer to the new challenges posed by AI considering sensitive data, such as personal data?

As we have witnessed, since 2021, by the legislative process of the EU AI Act (*vide*, [6]) it is not possible to legislate at the same speed as AI development. It is a competition where humans and machine cannot compete as equals. It is a race that humans would lose even before they began. So, the answer relies on applying old instruments to new challenges. In other words, looking for the answers for new problems in old solutions.

Rethinking the proportionality principle in the light of these new challenges and assuring that humans have, definitely, the last word, is one way of doing it. Therefore, we have analysed the proportionality principle within the scope of the art. 5.°, n. ° 1, par. h) of the EU AI Act [6] and the art. 9.° of the GDPR [26].

These articles address AI prohibited practices within the treatment of special categories of personal data such as biometric data. The need for a combined interpretation of these articles result, itself, from the actual reading of the EU AI Act art.° 5.°/1/h, in fine⁹ [6]. The reading, analysis and interpretation of these articles combined allowed us

⁹ EU AI Act (European Parliament and Council, 2024, p. 51–52): “Article 5 – Prohibited AI practices. 1. The following AI practices shall be prohibited: [...] (h) the use of ‘real-time’ remote biometric identification systems in publicly accessible spaces for the purposes of law enforcement, unless and in so far as such use is strictly necessary for one of the following objectives: (i) the targeted search for specific victims of abduction, trafficking in human beings or sexual exploitation of human beings, as well as the search for missing persons; (ii) the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or a genuine and present or genuine and foreseeable threat of a terrorist attack; (iii) the localisation or identification of a person suspected of having committed a criminal offence,

to identify, isolate and measure the frequency of expressions that clearly denounce the presence of the proportionality principle. For instance, the Eu AI Act, art. ° 5.º/1/h, uses expressions as “strictly necessary”. And the GDPR, art. ° 9.º, n. ° 2, par. a) to j)¹⁰, uses

for the purpose of conducting a criminal investigation or prosecution or executing a criminal penalty for offences referred to in Annex II and punishable in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least four years. Point (h) of the first subparagraph is without prejudice to Article 9 of Regulation (EU) 2016/679 for the processing of biometric data for purposes other than law enforcement.

- ¹⁰ Article 9 GDPR – Processing of special categories of personal data. 1. Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited. 2. Paragraph 1 shall not apply if one of the following applies: (a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject; (b) processing is necessary for the purposes of carrying out the obligations and exercising specific rights of the controller or of the data subject in the field of employment and social security and social protection law in so far as it is authorised by Union or Member State law or a collective agreement pursuant to Member State law providing for appropriate safeguards for the fundamental rights and the interests of the data subject; (c) processing is necessary to protect the vital interests of the data subject or of another natural person where the data subject is physically or legally incapable of giving consent; (d) processing is carried out in the course of its legitimate activities with appropriate safeguards by a foundation, association or any other not-for-profit body with a political, philosophical, religious or trade union aim and on condition that the processing relates solely to the members or to former members of the body or to persons who have regular contact with it in connection with its purposes and that the personal data are not disclosed outside that body without the consent of the data subjects; (e) processing relates to personal data which are manifestly made public by the data subject; (f) processing is necessary for the establishment, exercise or defence of legal claims or whenever courts are acting in their judicial capacity; (g) processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject; (h) processing is necessary for the purposes of preventive or occupational medicine, for the assessment of the working capacity of the employee, medical diagnosis, the provision of health or social care or treatment or the management of health or social care systems and services on the basis of Union or Member State law or pursuant to contract with a health professional and subject to the conditions and safeguards referred to in paragraph 3; (i) processing is necessary for reasons of public interest in the area of public health, such as protecting against serious cross-border threats to health or ensuring high standards of quality and safety of health care and of medicinal products or medical devices, on the basis of Union or Member State law which provides for suitable and specific measures to safeguard the rights and freedoms of the data subject, in particular professional secrecy; (j) processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject. 3. Personal

expressions as: par. b), c), d), f), g), h), i) “necessary” and “appropriate safeguards”; par. g), j), i) “shall be proportionate to the aim pursued” and “suitable and specific measures”.

The text analysis and interpretation allowed us to extract the conclusions as follows: firstly, the high frequency usage of the expressions that refer to the proportionality principle intend to clearly mention and apply the proportionality principle within this scope; secondly, the usage of expressions that specifically refer to the several sub-dimensions of the proportionality principle evidence that the sub-dimensions of adequacy, necessity and proportionality in a strict sense (i.e., pondering the balance between advantages and disadvantages of measures adopted towards the ends [25, p. 266–273], must be considered; finally, given the exceptions considered in art.º 5.º, of the EU AI Act combined with the art.º 9.º of the GDPR and the proliferation of expressions that refer to the proportionality principle in these articles, suggests that the proportionality principle core significance may be refer to as the last frontier for the protection of sensitive personal data. Therefore, the proportionality principle may potentially contribute to the actual and ultima ratio protection of sensitive data such as biometric data, neuro-rights and act as a protection barrier from mental hacking as long as the usage of data collected by neuroimage techniques observe the demands of the proportionality principle. At least, as far as the EU AI Act is concerned. Nevertheless, the last word still remains in the human domain as it is what seems to result from the reading of the art.º 9.º, n.º 3 and n.º 4 of the GDPR.

So, as scary as it maybe, AI is not scary at all, as some may say. Human fear towards AI systems usually results from ignorance, unawareness or misunderstanding about the functioning of AI systems which consequently generates wrongful human expectations towards AI systems capacities (e.g., AI systems don't have conscious, emotions, sensibility and sensitivity) [9]. So, what may be scary is what lies beneath, i.e., the human choices that lead us towards one path or another. So, maybe one of the cornerstones rely on the process of making educated human choices that result from implementing AI literacy measures by governments, since early to elder ages, that enable human beings with the necessary knowledge and instruments to think, create and use AI systems that observe and respect human rights.

5 Conclusions

Thus, finally, the real challenge related to the misuse or abuse of AI does not lie solely in the technical regulation of systems, but rather in the delicate pursuit of a sustainable equilibrium between the protection of Human Rights, the praxis of the proportionality principle, and the effective regulation of AI systems. This is especially critical in the context of systems—such as Generative AI—that involve the collection, processing, and

data referred to in paragraph 1 may be processed for the purposes referred to in point (h) of paragraph 2 when those data are processed by or under the responsibility of a professional subject to the obligation of professional secrecy under Union or Member State law or rules established by national competent bodies or by another person also subject to an obligation of secrecy under Union or Member State law or rules established by national competent bodies. 4. Member States may maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health.

storage of sensitive data, including facial and biometric information as well as neural data. These categories of data raise complex legal, ethical, and societal concerns due to their intrinsic link to individual identity, bodily integrity, and cognitive autonomy.

Consequently, the regulatory debate must encompass a wide spectrum of interrelated factors, such as ethical principles, algorithmic bias, data privacy and ownership, copyright regimes, and, crucially, digital literacy and education. These are not peripheral considerations but central pillars of a normative infrastructure that must be jointly constructed and upheld by algorithm stakeholders (including mathematicians, engineers, and technology providers), corporate actors, public institutions, and civil society at large [1]. In this regard, both state and non-state actors share the responsibility for fostering environments in which AI technologies can evolve within frameworks of accountability and trust.

Preliminary research findings indicate that AI literacy initiatives—whether driven by private or public sectors—play a vital role in reinforcing this fragile balance between regulatory compliance and Human-AI sustainable development [27], pp. 23–25]. By equipping individuals with the cognitive tools necessary to understand, critically engage with, and responsibly use AI systems, such initiatives contribute not only to individual empowerment but also to collective democratic resilience in the digital age.

At the policy level, the European Parliament has acknowledged this multidimensional challenge by supporting comprehensive strategies aimed at strengthening digital capacities, enhancing innovation, and ensuring that technological progress is aligned with the European Union’s values and commitments to fundamental rights. As highlighted in the institution’s 2024 communication, the European Parliament plays a key role in shaping policies that aim to expand digital technologies, promote sustainability through the green transition, foster digital inclusion and training for workers, and digitize public services—while rigorously safeguarding basic rights and democratic principles [28].

National digital strategies developed by EU Member States reflect these objectives. In Portugal, for instance, the *Resolução do Conselho de Ministros* n.º 207/2024, issued on 30 December 2024, addresses the complex challenges posed by the digital era and underscores the urgency of innovation and digital transformation [29], p. 1]. This led to the approval and implementation of the *Estratégia Digital Nacional* (EDN – National Digital Strategy), which embraces a multisectoral and multi-stakeholder approach, engaging the public and private sectors, civil society, and academia [30].

The EDN structures around four core dimensions—People, State, Business, and Infrastructures—with digital empowerment as a foundational objective, particularly within the “People” axis. The promotion of digital literacy, especially in relation to AI, is a central component of the strategy [29], p. 6]. Accordingly, national digital strategies that integrate AI education across all sectors and educational levels—such as the inclusion of mandatory AI curricula from early education through secondary school, lifelong learning programs targeting vulnerable populations, and tailored AI training for professionals in high-impact sectors like education, health, justice, and security—emerge as essential tools in achieving a balanced model of Human-AI co-evolution.

Ultimately, these educational and policy efforts converge toward the construction of a digitally literate and ethically conscious society, one in which the use of AI systems

is informed, deliberate, and always subordinate to human judgment. Ensuring that the final word rests with an educated and autonomous human decision-maker is the only path through which AI development can genuinely support and uphold Human Rights in practice.

References

1. Consultation Paper on AI Regulation: Emerging Approaches across the World. UNESCO. Paris, France (2024). <https://www.hkdca.com/wp-content/uploads/2024/08/consultation-paper-ai-regulation-unesco.pdf>
2. Saraiva, A.: Ethics and Responsibility in Artificial Intelligence: A Global Perspective and Portugal's AI Governance. Office for Strategy and Studies of the Ministry of Economy. Government of Portugal. (2024). https://www.gee.gov.pt/RePEc/WorkingPapers/GEE_PAPERS_186.pdf
3. Gellers, J.C.: Artificial Intelligence, Animal and Environmental Law. Routledge Taylor & Francis Group, New York (2021)
4. Rogerson, S.: Ethical Digital Technology in Practice. CRC Press, Taylor & Francis Group, London (2023)
5. White House's Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI. <https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>. Last accessed 1 May 2025
6. The Artificial Intelligence Act – Regulation (EU) 2024/1689. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>. Last accessed 1 May 2025
7. Gstrein, O.J., Haleem, N., Zwitter, A.: General-purpose AI regulation and the European Union AI Act. Internet Policy Rev. (2024). <https://doi.org/10.14763/2024.3.1790>
8. Innerarity, D.: Artificial Intelligence and Democracy. UNESCO and CLACSO. Paris, France. (2024). <https://unesdoc.unesco.org/ark:/48223/pf0000389736>
9. Mollick, E.: Co-inteligência: viver e trabalhar com IA. Trad. Luís Gonçalves. 1.^a ed. Porto Editora, Porto (2024)
10. Aguiar, F.: Direitos Humanos e Inteligência Artificial: principais dimensões jurídicas, éticas, sociais e culturais, no contexto europeu e transnacional. In: NETO, Carlos et al. (coord.) – Mentres digitais: do zero ao infinito. Homenagem à Professora Fernanda Duarte. Pref. Guilherme Gama. Editora GZ, Rio de Janeiro, pp. 1–36 (2024)
11. Damásio, A.: Sentir & Saber: a caminho da consciência. 1.^a ed. Temas e Debates, Círculo de Leitores, Lisboa (2020)
12. Sigman, M.: O poder das palavras: a arte de conversar. Trad. Ana Mendes. Temas e Debates. 1.^a ed. Bertrand Editora, Lisboa (2023)
13. Rovelli, C.: A ordem do tempo. Trad. Silvana Cobucci. 1.^aed., 4.^a Republicação. Penguin Random House, Lisboa (2022)
14. Brazil, Câmara dos Deputados, Projeto de Lei 522/22, Portal da Câmara dos Deputados. <https://www.camara.leg.br/propostas-legislativas/2317524>. Last accessed 1 May 2025
15. Chile. MinCiencia Ministerio de Ciencia, Tecnología, Conocimiento e Innovación – Ley 21383/2021. Biblioteca del Congreso Nacional de Chile. <https://www.bcn.cl/leychile/navegar?idNorma=1166983>. Last accessed 1 May 2025
16. Martinho, A.: Digitalização, automação e Inteligência Artificial nos Tribunais Portugueses. REVISTA (03 Cultural), pp. 167–178. Supremo Tribunal de Justiça, Lisboa (2023)
17. Liu, J.Z., Li, X.: How do judges use large language models? Evidence from Shenzhen. J. Legal Anal. **16**, 235–259 (2025). <https://doi.org/10.1093/jla/laae009>

18. American Psychology Association (APA): Dictionary of Psychology, Suggestibility. <https://dictionary.apa.org/suggestibility>. Last accessed 1 May 2025
19. Nunes, B.: *Memória de ferro: aprenda a cuidar da sua memória nas diferentes fases da vida*. Presença, Barcarena (2023)
20. Rovelli, C.: *Sete breves lições de física*. Trad. Vasco Gato. 1.^a ed., 3.^a Republicação. Penguin Random House, Lisboa (2022)
21. Brazil, Senado, Projeto de lei 2.338/2023. <https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>. Last accessed 14 Apr 2025
22. Brazil, Câmara dos Deputados, Projeto de lei 2174/2023. <https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2358605>. Last accessed 14 Apr 2025
23. OECD Framework for the Classification of AI Systems. https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html. Last accessed 1 May 2025
24. Ferrari, I.: *Discriminação algorítmica e Poder Judiciário: limites à adoção de decisões algorítmicas no Judiciário brasileiro*. 1.^a Ed. Emais, Florianópolis (2023)
25. Canotilho, J.G.: *Direito Constitucional e Teoria da Constituição*, 7.^a Ed. Almedina, Coimbra (2023)
26. Regulation (EU) 2016/679 of the European Parliament, Directive 95/46/EC (General Data Protection Regulation). <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>. Last accessed 14 Apr 2025
27. Mazzi, F., Taddeo, M., Floridi, L.: AI in support of the SDGs: six recurring challenges and related opportunities identified through use cases. In: Mazzi, F., Floridi, L. (eds.) *The Ethics of Artificial Intelligence for the Sustainable Development Goals*. Philosophical Studies Series, vol. 152. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-21147-8_2
28. European Parliament, Shaping the digital transformation: EU strategy explained. <https://www.europarl.europa.eu/topics/en/article/20210414STO02010/shaping-the-digital-transformation-eu-strategy-explained>. Last accessed 14 Apr 2015
29. Portugal, Presidência do Conselho de Ministros, Resolução do Conselho de Ministros n.º 207/2024. <https://diariodarepublica.pt/dr/detalhe/resolucao-conselho-ministros/207-2024-901536081>. Last accessed 1 May 2025
30. Portugal, Governo, Estratégia Digital Nacional (EDN), <https://digital.gov.pt/documentos/estrategia-digital-nacional>. Last accessed 1 May 2025



Rethinking Educational Assessment in the Age of Artificial Intelligence

Mario Arias-Oliva^{1,2} , Antonio Pérez-Portabella^{1,2} , Manuel Ollé Sesé¹ ,
and Jorge de Andrés Sánchez² 

¹ Complutense University of Madrid, Madrid, Spain
mario.arias@ucm.es

² Universitat Rovira i Virgília, Tarragona, Spain

Abstract. As artificial intelligence (AI) is rapidly transforming the landscape of education, its influence on assessment practices has become both transformative and controversial. This paper analyzes the potential benefits and concerns of AI-driven educational assessment, with particular attention to the ethical, pedagogical, and social implications of its widespread adoption. Emerging technologies such as Large Language Models (LLM) and machine learning algorithms are enabling new efficient forms of feedback, personalization, and scalability that were unimaginable a few years ago. Nevertheless, many unresolved challenges should be considered, such as algorithmic bias, academic integrity, or legal issues. It is necessary an urgent reframe of AI based assessment through a more ethical and human-centered perspective. By foregrounding values such as fairness, transparency, AI literacy, or inclusivity among other aspects, this work outlines a roadmap for integrating AI tools in ways that enrich and improve assessment experiences, avoiding at the same time drawbacks. The paper proposes ideas to design future AI assessment systems that are both innovative and deeply responsible for students, educators, technologists, and policymakers, finding an ethical perspective for future deployment of smart assessment technologies.

Keywords: AI Assessment · AI Education · Ethics of AI Assessment · Smart Assessment

1 Introduction

Over the past decade, artificial intelligence (AI) has significantly reshaped the landscape of educational assessment. What began as a tool for automating simple tasks, like grading multiple-choice exams or organizing test logistics, has evolved into something far more dynamic and capable. With the rise of powerful technologies such as Large Language Models, AI systems are now helping both educators and students, in innovative more meaningful ways, by offering timely, consistent, and often remarkably accurate feedback on student work. These tools can evaluate a wide range of responses, from essays and short answers to spoken presentations, doing so at a scale and speed that would be nearly impossible for human lecturers to match alone. By taking over repetitive grading

duties, AI allows educators to focus more on what really matters: supporting students individually, refining their teaching strategies, and encouraging deeper thinking and learning in the classroom.

Current innovations in Machine Learning (ML), Natural Language Processing (NLP) and the development of Large Language Models (LLMs), have greatly expanded the potential of AI in learning, including assessment methods, process and tools. Nowadays, innovative artificial intelligence technologies are proving effective not only in assessing open-ended responses with remarkable accuracy, but also in generating personalized feedback, developing assessment items and performing predictive analysis of student performance. Through adaptive feedback loops, AI can now respond to individual learning patterns, offer encouragement, clarify misunderstandings, and even suggest tailored next steps. This shift marks a movement away from a rigid, generic, undifferentiated and non-customized assessment model toward a more responsive, student-centered approach that has the potential to transform how we understand and support learning in real time. For example, NLP facilitates the automated assessment of essays and provides real-time feedback on written work [1, 2], while ML algorithms can identify students at academic risk based on performance indicators [3]. Generative AI applications, such as ChatGPT, Gemini, Perplexity or Claude, among many others, are currently being investigated for functions such as the automated generation of assessment items [4]. Furthermore, these tools have evolved beyond passive instruments to become interactive systems that engage students through adaptive feedback mechanisms, enabling more individualized educational experiences. This means educators are no longer the sole architects of learning materials. AI is becoming a creative partner in the assessment process for lecturers and students.

However, while the integration of AI into learning assessment offers significant advantages, it also raises critical ethical and social concerns. Issues such as algorithmic accountability, transparency and explicability; evaluative fairness; assessment validity; bias; educational equity; institutional accountability; or academic integrity, among others, must be carefully considered to ensure the ethical implementation of AI-driven assessment systems. Addressing these concerns requires the adoption of methodological frameworks that actively involve human oversight to support the responsible development and integration of smart technologies [3, 5]. This paper focuses on analyzing and conceptually framing the ethical implications associated with the use of AI in educational assessment.

2 Methodology

This conceptual paper seeks to offer a thoughtful and well-grounded perspective on the ethical and pedagogical challenges that artificial intelligence introduces into the realm of educational assessment. To support this objective, we carried out a comprehensive review of relevant scholarly literature, with the aim of identifying the most relevant and widely discussed dimensions in current academic and professional debates in the field of AI assessment. Our research involved systematic searches across major academic databases—such as Web of Science (WoS), Scopus, Google Scholar, JSTOR, and Dialnet—giving priority to peer-reviewed journal articles and recent contributions

that explore the intersection of AI ethics, educational technologies, and evolving models of assessment. Other external sources were included as well in order to include practitioners' perspectives.

The selection of academic sources for this study followed the criteria proposed by Paul and Criado in their 2020 work [6], emphasizing thematic relevance over metrics such as journal impact factor or citation count. For the analysis of the retrieved literature, we employed content analysis methodology, understood as a systematic, objective, and qualitative approach to the interpretation of textual material [7]. In this case, we focused on documents addressing the ethical, pedagogical, and institutional dimensions of AI-driven assessment. Following [8] the sampling units were the selected publications, from which we identified and categorized key analytical themes. These were then contextualized through a cross-analysis of sources, supported by complementary searches to refine and triangulate core concepts [9]. This contextual approach allows for the integration of our findings into the broader research landscape, helping to detect prevailing trends and uncover potential gaps in the current literature [10].

3 Literature Review on the Ethical Challenges on the Integration of AI into Assessment

In this section we will introduce related concepts about the integration of AI into assessment, describing what could be their potentially drawbacks and negative consequences. Following the methodology outlined in the previous section, we have structured the literature review around three core areas. The first (Sect. 3.1) focuses on Transparency, Explainability (XAI), and Accountability, which are grouped together due to their conceptual interconnections. The second area (Sect. 3.2) brings together related concepts: Bias, Fairness, Equity, and Legal Aspects. Finally, and identified as particularly significant in our analysis, we address the issue of Academic Integrity. The purpose of this literature review is to examine the key concepts related to the ethical and problematic dimensions of applying AI in educational assessment.

3.1 Transparency, Explainability (XAI) and Accountability in AI Assessment

Transparency in artificial intelligence refers to the clarity and openness regarding how AI systems operate, including their design, decision-making processes, and limitations. It enables individuals to understand how AI systems make decisions that affect their lives, fostering trust and accountability in AI governance [11]. Explainable Artificial Intelligence (XAI) refers to the development of AI systems that enhance transparency by providing clear insights into how outputs and decisions are generated [12]. Accountability is a complex concept, and its application becomes even more challenging when is applied to emerging technologies such as artificial intelligence. In the context of AI, accountability refers to the responsibility of developers, deployers, and users to justify, take ownership and assume consequences of the outcomes produced by artificial intelligence systems. This involves ensuring that these systems operate transparently, comply with ethical standards, and are subject to effective oversight mechanisms that allow the evaluation and remediation of their societal and individual negative impacts. Effective

AI accountability frameworks involve clear delineation of responsibilities, the ability to audit and explain AI decisions, and mechanisms for addressing any adverse effects resulting from AI deployment [13].

AI systems often function as opaque “black boxes”, in which inadequate interpretability undermines trust in AI-mediated decisions. In an assessment process, lack of accountability, transparency and explainability are significant barriers to building the required trust, affecting acceptance of artificial intelligence tools in assessment. Neither students nor educators should be expected to accept the outcomes of a black-box system without critically examining the mechanisms behind its decisions. In some cases, such decisions may carry serious consequences for individuals and institutions, making it essential to establish accountability processes that clearly define responsibilities. Therefore, both developers and users must advocate for these principles (accountability, transparency, and explainability) to support the responsible development and meaningful integration of AI in educational assessment.

In education, transparency encompasses the degree of disclosure provided by AI developers and educational institutions about how these systems work, including architectural design parameters, data input specifications, algorithmic decision-making frameworks, and inherent limitations. Explainability, on the other hand, refers to the intelligibility of AI-generated results to non-specialists, such as students, educators, and policymakers. These emerging technological tools are used each time in more stages and processes of assessment, such as assessment test design, grading, pedagogical feedback, or admission decisions, becoming fundamental understanding their operating mechanisms in order to ensure ethical integrity and institutional trust. Explainability is a critical dimension that should be integrated for an ethical use of AI in assessment [14]. Within educational assessment paradigms, when an automated assessment algorithm assigns a suboptimal score to a student’s writing, both the student and the lecturer must be able to discern the basis for the assessment, whether it is based on grammatical construction, rhetorical coherence, or content adequacy. Awareness of all these factors is fundamental for the proper development of AI-based assessment. Without strong frameworks for transparency and explainability, there is a real risk of unfair outcomes, especially when there are no clear ways to question or raise objections against mistakes or bias in algorithmic decisions.

3.2 Bias, Fairness, Equity and Legal Aspects

Bias in artificial intelligence refers to systematic errors in decision-making processes that lead to unfair outcomes, often disadvantaging certain individuals or groups. In artificial intelligence systems, bias can emerge from various sources, including unrepresentative training data, algorithmic assumptions, or user interactions. For example, biased datasets can result in models that perform well for majority populations while consistently underperforming for marginalized communities, leading to discriminatory outcomes in high-stakes contexts such as healthcare, hiring, and education. These biases are not merely technical flaws; they reflect and reinforce existing social inequalities [15]. Bias in educational AI systems is related to systematic errors that result in unfair or discriminatory outcomes for certain student populations. In education, such bias often stems from datasets that fail to represent the diversity of learners in terms of race, gender,

socioeconomic background, or learning styles. For example, machine learning models used in assessment may be trained on data primarily sourced from majority populations, thereby marginalizing underrepresented groups. These systems can inadvertently reinforce existing disparities, leading to lower performance evaluations or misclassification of students' abilities [16]. Addressing bias in educational AI is not only a technical issue but also a deeply ethical concern that demands critical reflection on whose knowledge is valued and how learning is defined.

Fairness in AI is a broader concept. While bias often arises unintentionally, fairness involves deliberate efforts to ensure that AI systems do not favor or disadvantage any individual or group based on protected characteristics like race, gender, or socioeconomic status. Different fairness frameworks have been proposed, such as group fairness, which ensures equitable outcomes across demographic groups, and individual fairness, which requires that similar individuals be treated similarly by the system. Counterfactual fairness, a newer model, seeks to guarantee that an AI decision would remain the same even if sensitive attributes were changed [17]. These frameworks illustrate that fairness is context-dependent and ethically grounded, requiring transparent and accountable development practices. Fairness in AI-based educational assessment goes beyond simply removing bias, going beyond with the aim of creating systems that uphold principles of equity and justice. This means designing tools that not only avoid harming students but also recognize and respond to their diverse backgrounds and needs. Fairness can take different forms. Group fairness, for example, focuses on making sure that students from different demographic groups are treated equitably and achieve similar outcomes. Individual fairness aims to ensure that learners with similar characteristics receive similar treatment. In a classroom setting, this could mean that an AI system offering personalized feedback must adjust to each student's context rather than applying a single standard for everyone [18]. Another previously cited concept, counterfactual fairness, asks whether the system's decisions would stay the same even if a student's race, gender, or other protected attributes were different [19]. Taken together, these ideas show that fairness in education is not about rigid equality. Fairness requires thoughtful, context-sensitive design that adapts to the real-world complexity of learners and classrooms.

Equity, closely linked to fairness, emphasizes the need to recognize and address historical and structural disadvantages that affect different groups. Equity in AI-based educational assessment goes beyond fairness by actively recognizing and addressing the deep-rooted inequalities that have shaped students' learning opportunities. It calls for technologies that are thoughtfully designed to meet the diverse needs of all learners, especially those from historically marginalized communities. This means using training data that truly reflects that diversity, involving both teachers and students in the design process, and regularly reviewing how these systems perform to catch any unintended negative effects [17]. For instance, an AI tool that consistently underestimates the potential of students from low-income schools. In that case, applying equity requires not just to adjust the model, but also to rethink how educators use and interpret its results. Another example could be designing assessment tools that are sensitive to linguistic or cultural variations among students or ensuring that datasets reflect diverse learning experiences.

As highlighted by the European Commission [20], for AI to be truly equitable in education, it must be transparent, accountable, and deeply grounded in human values, because these tools are not just making predictions; they are shaping real lives in real classrooms.

The integration of AI tools in educational assessment raises important legal concerns, particularly around data protection, discrimination, and accountability. For instance, the General Data Protection Regulation (GDPR) [21] in the European Union is especially relevant, as it prohibits solely automated decisions with significant effects, such as grading, unless explicit consent is obtained and human intervention is ensured. AI applications, like automated essay scoring, often operate on sensitive personal data, invoking the need for robust compliance mechanisms [22]. Other legal debate focuses on the risk of algorithmic bias, as AI models have been trained on historical data, they can replicate and reinforce structural inequalities. This may contravene anti-discrimination frameworks, particularly when underrepresented or marginalized student groups are adversely affected [24]. In cases where AI-driven decisions are found to be faulty or harmful, the question of accountability becomes complex. Responsibility can be distributed across developers, educational institutions, and policymakers, raising challenges in legal redress and governance [25]. These concerns illustrate the necessity for transparent and human mediated systems, and a strong regulatory framework to uphold legal rights and educational equity.

Determining responsibility in cases where AI systems generate errors or produce unintended consequences in the context of assessment is an important ethical challenge. As AI tools are progressively integrated into educational assessment methodologies, establishing explicit accountability structures becomes critically important. These frameworks require the precise identification of the distribution of responsibility among multiple stakeholders, including technology developers, the educational institutions that implement them, and professional educators, when AI systems misjudge academic work, disseminate misleading formative feedback, or perpetuate existing systemic biases [26].

3.3 Academic Integrity Challenges

Academic integrity involves not only the avoidance of dishonest behaviors such as plagiarism or fraud, but also the development of learning environments where all participants (students, educators, and administrators) share a collective commitment to act ethically and transparently. This broader understanding highlights integrity as a shared responsibility that sustains the trust and credibility, essential factors to develop a meaningful learning and fair assessment. Within the context of assessment, academic integrity can be understood as a fundamental ethical principle that ensures learning evaluations honestly and fairly reflect students' knowledge, skills, and effort [27].

Academic dishonesty is not a new phenomenon—it has been part of educational settings long before the development of artificial intelligence. Practices such as contract cheating, identity fraud, and cheating during exams have been persistent challenges in higher education. What AI has done is introduce new modalities and tools that may facilitate such behaviors and develop new ones, but the underlying ethical concerns remain rooted in longstanding issues of academic integrity [28]. An important factor to consider is the facilitation of academic misconduct through AI-generated content. AI tools can be used to produce assignments that students may submit as their own, thereby

undermining the authenticity of academic work. Although misconduct involving AI is a newer form, it reflects long-standing patterns of academic dishonesty. This means that existing strategies to prevent cheating need to be updated to respond to these emerging challenges [29]. Furthermore, the widespread availability of AI-generated content complicates the detection of plagiarism and other forms of academic dishonesty. Traditional plagiarism detection tools may not be equipped to identify AI-generated text, making it more difficult for educators to uphold academic standards.

AI-generated text detection tools, such as Turnitin's AI detection software, have been criticized for their unreliability and potential biases. These tools can produce false positives, unfairly targeting students who write in a polished or formal style and may disproportionately affect non-native English speakers or those with learning differences [30].

In response to these challenges, educational institutions are encouraged to revise their academic integrity policies and incorporate AI literacy into their curricula. By educating students about the ethical use of AI and the importance of maintaining academic honesty, institutions can foster a culture that supports the values of integrity even in the face of rapidly advancing technologies. Strategies to address academic dishonesty in online assessments should be developed, such as implementing clear honor codes, fostering an institutional culture that genuinely values integrity, and ensuring that both students and educators are well-informed about academic policies and the consequences of misconduct. Another important factor to consider is the design of assessment tools that are authentic and diverse. These assessments should require students to complete complex tasks that cannot be easily generated with a single prompt in a Large Language Model. By encouraging critical thinking and promoting high levels of student engagement, such assessments make dishonest behavior significantly more difficult to carry out. This approach represents a shift away from punitive measures toward more preventive and educational strategies, grounded in mutual trust and a sense of shared responsibility [31].

From the lecturer's perspective, there are as well important concerns about the use of AI in assessment. Lecturer tasks, such as the automated grading of essays, the creation of exams, or the design of case studies for student assessment, reveal another dimension of academic integrity, the one that concerns educators themselves. These practices highlight that integrity is not solely a student responsibility; it also involves the ethical use of technology by faculty in ways that preserve the fairness, transparency, and educational value of the assessment process. The growing use of AI tools in grading is reshaping not just our workflows, but also our educational values. The promise of quicker feedback and reduced workloads is undeniably attractive, especially for faculty managing heavy teaching loads. Relying on AI for grading, particularly when motivated by efficiency, may unintentionally compromise key principles such as academic integrity, student privacy, and the depth of evaluative feedback [32]. It is necessary to think carefully about how these technologies align with the core values of their institutions. Rather than simply adopting AI tools for convenience, it is necessary a deliberate, transparent, and ethically informed approach, one that keeps human judgment at the heart of assessment. Institutional guidance and professional development reflect is a must to ensure a proper integration of AI into our practices.

4 Discussion

The integration of AI technologies into educational assessment represents a meaningful shift away from traditional evaluation practices. It calls for a careful rethinking of the skills and knowledge that educators and other stakeholders will need to remain effective in the future [33]. While AI offers important opportunities to personalize learning and promote greater inclusion, particularly for diverse learners and students with special educational needs, it also introduces a set of complex challenges that reshape how we think about assessment and its role in education.

Fostering AI literacy has become a crucial priority to ensure that emerging technologies are used appropriately and responsibly in educational contexts. This form of literacy goes beyond understanding how algorithms work; it also involves being critically aware of broader social issues, such as algorithmic bias, data privacy, and fair access to technology. AI literacy empowers both students and educators with the required knowledge and skills to engage with AI-generated content in an informed and reflective manner. Acquiring the necessary literacy in the use of AI tools enables students to engage critically with AI-generated content, ensuring they are not passive recipients but informed participants in AI-mediated educational environments. This literacy fosters active engagement, allowing learners to question, interpret, and meaningfully apply information in their learning processes. One key aspect of this literacy is the mastery of prompt engineering, a specialized skill that enables users to craft precise and purposeful prompts to guide Large Language Models (LLMs) in producing relevant, accurate, and pedagogically useful outputs. Mastery of prompt engineering is required, a specialized skill that enables users to generate accurate, contextually relevant and pedagogically valuable learning. This skill strengthens students' ability to collaborate with AI tools, turning assessment into a more interactive and exploratory learning experience. In addition, critical thinking becomes even more important in AI-supported education, as students need to evaluate, question, and verify the information generated by these systems. Consequently, the future evolution of educational assessment depends on the ability of students and educators to interact meaningfully with AI technologies, interpret their results with appropriate skepticism, and integrate these interpretations into reflective learning practices that improve educational outcomes [34].

Accountability is another keystone of an ethical and appropriate development of AI-Assessment. Without clear accountability frameworks, both students and educators can be left exposed to the consequences of flawed algorithmic decisions, with few options to raise concerns or correct mistakes. Responsibility must be a core principle in the ethical use of AI in education, making sure that human oversight stays central to how assessment systems are designed and used [11]. Understanding and addressing of bias is crucial to building systems that serve all users equitably. With regard legal challenges to develop a trustworthy AI, European Comision propose 3 guidelines: legality, ethics, and robustness. First, AI systems must comply with all relevant laws and regulations, ensuring legal accountability. Second, they should reflect ethical principles and values, safeguarding human dignity, autonomy, and fairness. Third, they must be robust, both technically and socially, to minimize the risk of unintended har, even when deployed with the best intentions. While each component is crucial on its own, true trustworthiness in AI emerges only when all three dimensions work together in balance [20].

With respect to other approaches to successful integration, it is worth highlighting that strategies focusing on banning or limiting access to generative AI tools have largely fallen short, and punitive measures from administrators often overlook the valuable educational opportunities these technologies can provide. As a result, the reconceptualization and reform of assessment practices have become an imperative priority in contemporary educational discourse during this transformative era of AI [34].

Education in general, and assessment particularly, require the implementation of innovative methodological frameworks that include humans in the process to ensure the responsible deployment and integration of technology [3, 5]. Challenges are not the mere use of Artificial Intelligence (AI) tools that constitutes plagiarism or a breach of academic integrity. Rather, what truly matters is the level of transparency with which students disclose their use of these technologies. The ethical concern arises not from the technology itself, but from how it is integrated into the learning and assessment process. For instance, a student who uses an AI-based language model such as ChatGPT to brainstorm ideas or rephrase a paragraph, and clearly indicates this in their submission, demonstrates an awareness of academic responsibility and upholds the principles of integrity. In contrast, submitting an AI-generated essay as one's own original work, without any acknowledgment, not only misrepresents authorship but also undermines the trust on which academic communities are built. This distinction underscores the importance of institutional policies that encourage students to be open about their use of AI, while also guiding them on how to engage with these tools ethically. As AI becomes increasingly integrated into educational environments, fostering a culture of honesty and reflective practice will be crucial to maintaining the credibility and fairness of academic evaluation [35].

5 Conclusions, Limitations and Future Studies

AI technologies hold significant promise in educational assessment, but their integration requires a careful, ethically grounded approach. These tools should not merely replace traditional practices but rather enhance them through intentional design that prioritizes transparency, fairness, and sustained human oversight. Fostering AI literacy among both educators and students is essential—not only to enable effective use but also to promote a critical understanding of the broader social and ethical implications of these technologies. Addressing potential risks such as algorithmic bias, opacity, and AI-facilitated academic dishonesty calls for robust institutional support. This includes rethinking assessment models and integrity policies, moving beyond punitive or prohibitive responses. AI in education is not a neutral force; its deployment must reflect and reinforce human values. Ensuring fairness, transparency, and ethical oversight is crucial to harnessing its benefits while protecting the integrity of educational processes.

Regarding the limitations and future research of this study, it constitutes an exploratory investigation aimed at identifying some of the critical areas related to the responsible implementation of artificial intelligence in assessment processes. Further research will be necessary, among which we preliminarily propose the development of qualitative studies, using semi-structured interviews with key stakeholders, including faculty members, students, and institutional leaders. This method enables a rich and

detailed exploration of diverse perspectives, real-world experiences, and determine the situation of factors described in this paper and detect other emerging concerns related to AI-driven assessment tools. Semi-structured interviews are especially appropriate for this inquiry, as they offer a balance between guided thematic exploration and the flexibility needed to capture the variety of viewpoints present across higher education institutions [36].

Acknowledgments. This research was funded by the Complutense University Research Project titled “Impacts of Artificial Intelligence on the Tourism Industry from a Social Science Perspective” (project number 506-2024), and the Telefónica Chair on Smart Cities at Universitat Rovira i Virgili and Universitat de Barcelona (project number 42.DB.00.18.00).

References

1. Hockly, N.: Automated writing evaluation. *ELT J.* **73**(1), 82–88 (2019). <https://doi.org/10.1093/elt/ccy044>
2. Ke, Z., Ng, V.: Automated essay scoring: a survey of the state of the art. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), pp. 6300–6308 (2019). <https://doi.org/10.24963/ijcai.2019/879>
3. Bulut, O., Beiting-Parrish, M.: The rise of artificial intelligence in educational measurement: opportunities and ethical challenges. *Chinese/English J. Educ. Measur. Evaluat.* (2024). <https://doi.org/10.59863/MIQL7785>
4. Lohr, D., Berges, M., Chugh, A., Kohlhase, M., Müller, D.: Leveraging large language models to generate course-specific semantically annotated learning Objects. *J. Comput. Assist. Learn.* **41**, e13101 (2025). <https://doi.org/10.1111/jcal.13101>
5. Perkins, M., Furze, L., Roe, J., MacVaugh, J.: The AI Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment (2023). arXiv. <https://arxiv.org/abs/2312.07086>
6. Paul, J., Criado, A.R.: The art of writing literature review: What do we know and what do we need to know? *Int. Bus. Rev.* **29**(4), 101717 (2020). <https://doi.org/10.1016/j.ibusrev.2020.101717>
7. Berelson, B.: *Content Analysis in Communication Research*. Free Press, Glencoe, IL (1952)
8. Krippendorff, K.: *Content Analysis: An Introduction to its Methodology*, 2nd edn. Sage Publications, Thousand Oaks, CA (2004)
9. Flick, U.: Triangulation in data collection. In: Flick, U. (ed.) *The SAGE Handbook of Qualitative Data Collection*, pp. 527–544. SAGE Publications Ltd, 1 Oliver’s Yard, 55 City Road London EC1Y 1SP (2018). <https://doi.org/10.4135/9781526416070.n34>
10. Pedraza-Navarro, I., Sánchez-Serrano, S.: Análisis de las publicaciones presentes en WoS y Scopus. Posibilidades de búsqueda para evitar literatura fugitiva en las revisiones sistemáticas. *Revista Interuniversitaria de Investigación en Tecnología Educativa* (2022). <https://doi.org/10.6018/riite.548361>
11. Novelli, C., Taddeo, M., Floridi, L.: Transparency and accountability in AI systems. *Front. Hum. Dyn.* **5**, Article 1421273 (2023). <https://doi.org/10.3389/fhumd.2024.1421273>
12. Türkmen, G.: The review of studies on explainable artificial intelligence in educational research. *J. Educ. Comput. Res.* **63**(2), 277–310 (2024). <https://doi.org/10.1177/07356331241310915>
13. Novelli, C., Taddeo, M., Floridi, L.: Accountability in artificial intelligence: what it is and how it works. *AI Soc.* **39**, 1871–1882 (2024). <https://doi.org/10.1007/s00146-023-01635-y>

14. Doewes, A., Pechenizkiy, M.: Structural explanation of automated essay scoring. In: Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020), pp. 758–761. International Educational Data Mining Society (2020)
15. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**(6464), 447–453 (2019). <https://doi.org/10.1126/science.aax2342>
16. Baker, R.S., Hawn, A.: Algorithmic bias in education. *Int. J. Artif. Intell. Educ.* **32**, 1052–1092 (2021). <https://doi.org/10.1007/s40593-021-00285-9>
17. Ferrara, E.: Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. *Sci* **6**(1), 3 (2024). <https://doi.org/10.3390/sci6010003>
18. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226 (2012). <https://doi.org/10.1145/2090236.2090255>
19. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard J. Law Technol.* **31**(2), 841–887 (2018). <https://doi.org/10.2139/ssrn.3063289>
20. European Commission: Ethics guidelines for trustworthy AI. Publications Office (2019). <https://data.europa.eu/doi/https://doi.org/10.2759/346720>
21. European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official J. European Union **L119**, 1–88 (2016). <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
22. Huang, L.: Ethics of artificial intelligence in education: student privacy and data protection. *Sci. Insights Educ. Front.* **16**(2), 2577–2587 (2023). <https://doi.org/10.15354/sief.23.re202>
23. Gándara, D., Anahideh, H., Ison, M., Picchiarini, L.: Inside the black box: detecting and mitigating algorithmic bias across racialized groups in college student-success prediction. *AERA Open* **10**(1), 1–15 (2024). <https://doi.org/10.1177/23328584241258741>
24. Cheong, B.C.: Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Front. Human Dyn.* (2024). <https://doi.org/10.3389/fhumd.2024.1421273>
25. Williamson, B., Eynon, R.: Historical threads, missing links, and future directions in AI in education. *Learn. Media Technol.* **45**(3), 223–235 (2020). <https://doi.org/10.1080/17439884.2020.1798995>
26. Bretag, T. (ed.): Handbook of Academic Integrity. Springer, Singapore (2016). <https://doi.org/10.1007/978-981-287-098-8>
27. Morris, E.J.: Academic integrity matters: five considerations for addressing contract cheating. *Int. J. Educ. Integrity* (2018). <https://doi.org/10.1007/s40979-018-0038-5>
28. Birks, D., Clare, J.: Linking artificial intelligence facilitated academic misconduct to existing prevention frameworks. *Int. J. Educ. Integr.* (2023). <https://doi.org/10.1007/s40979-023-00142-3>
29. Coldwell, W.: ‘I received a first but it felt tainted and undeserved’: Inside the university AI cheating crisis. *The Guardian* (2024). <https://www.theguardian.com/technology/2024/dec/15/i-received-a-first-but-it-felt-tainted-and-undeserved-inside-the-university-ai-cheating-crisis>
30. Holden, O.L., Norris, M.E., Kuhlmeier, V.A.: Academic integrity in online assessment: a research review. *Front. Educ.* **6**, 639814 (2021). <https://doi.org/10.3389/feduc.2021.639814>
31. Kumar, R.: Faculty members’ use of artificial intelligence to grade student papers: a case of implications. *Int. J. Educ. Integr.* (2023). <https://doi.org/10.1007/s40979-023-00130-7>
32. Walter, Y.: Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education. *Int. J. Educ. Technol. High. Educ.* **21**, 15 (2024). <https://doi.org/10.1186/s41239-024-00448-3>

33. Yuan, B., Hu, J.: Generative AI as a Tool for Enhancing Reflective Learning in Students (2024). arXiv. <https://arxiv.org/abs/2412.02603>
34. Ta, R., West, D.M.: Should schools ban or integrate generative AI in the classroom? Brookings Institution (2023). <https://www.brookings.edu/articles/should-schools-ban-or-integrate-generative-ai-in-the-classroom/Brookings>
35. Perkins, M.: Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *J. Univ. Teach. Learn. Pract.* (2023). <https://doi.org/10.53761/1.20.02.07>
36. Kvale, S., Brinkmann, S.: *InterViews: Learning the Craft of Qualitative Research Interviewing*, 2nd edn. SAGE Publications (2009)

Systems: Tools and Technologies for Responsible AI



Evaluating the Role of Chatbots in Higher Education Based on Students' Experiences

Ewa Duda^(✉) , Ewelina Młynarczyk-Karabin ,
and Julia Więckiewicz-Modrzewska 

Maria Grzegorzewska University, Warsaw, Poland
{eduda, emlynarczyk, jwieckiewicz}@aps.edu.pl

Abstract. The integration of chatbots in higher education has emerged as a transformative tool to increase student engagement, streamline administrative processes and support personalised learning. This article examines the current uses, benefits and challenges of chatbot technology in academia, based on an empirical study of unique sample of 408 students at a pedagogical university, people who will shape the future education of others. The findings indicate that students are most likely to use ChatGPT, and to use it mainly in education-related areas, and to a lesser extent in areas of everyday life. In this respect, non-pedagogy students were more likely than pedagogy students to generate ready-made answers during exams. Despite students claiming to have a fairly good knowledge of chatbots, only a quarter had read articles or research on them. The study also explored ethical issues related to students' use of chatbots. Few respondents believe that their teachers or university have rules or guidelines for the responsible use of chatbots. Non-pedagogy students and pedagogy students with teaching qualifications rate this issue the lowest. Almost half of the respondents feel privacy and security concerns when using a chatbot. Seven out of ten believe that chatbots can be used to manipulate information or mislead students. Students also believe that chatbots could increase inequalities in education. Through a synthesis of existing literature and practical implementations, this paper provides insights for teachers, administrators and developers who want to use chatbot technology to foster innovation in university education.

Keywords: Artificial Intelligence · Chatbots · Future Teachers · Higher Education · Student Support

1 Introduction

In recent years, tools based on artificial intelligence (AI) have played an increasingly important role in various areas of life, including higher education. One of the most advanced applications of natural language processing (NLP) technologies is the use of large language models (LLMs), which power intelligent chatbots capable of generating text, answering questions, and supporting teaching and learning processes [1]. While ChatGPT is a leading example, many similar tools, such as Google's Gemini or Claude by Anthropic, are increasingly used in educational settings. Although the integration of such

technologies into university education remains controversial and raises significant ethical concerns, including issues of academic integrity, originality, privacy, and dependency, a growing number of students are engaging with LLM-based chatbots as part of their everyday learning routines.

On the one hand, chatbots can be a valuable tool to support students in their studies, helping them to write essays, translate content and analyse complex issues. It can also make education more accessible, support those with learning difficulties and facilitate the process of acquiring knowledge [2]. On the other hand, there are concerns about the reliability of information, academic integrity and the impact of technology on students' independent thinking. The potential for misuse of the tool, for example by generating academic content without adequate self-input, is also a major concern [3].

The aim of this study was to identify the challenges and opportunities related to the use of AI from the perspective of students preparing to become teachers. It focuses on areas where future teachers see potential applications of AI, as well as aspects they perceive as threats to the educational process and their professional development. Specifically, the study analyses how students of pedagogy use chatbots, such as ChatGPT, during their studies.

2 Large Language Models in Higher Education

The development of artificial intelligence (AI) technologies has had a significant impact on the transformation of higher education, introducing new teaching methods, personalising the learning process and automating teaching processes. Silva with colleagues [4] highlight that the use of AI can bring benefits such as tailoring learning materials to individual student needs, facilitating student-teacher interaction, and assisting in the assessment process. Their study suggests that the implementation of AI-based solutions not only improves teaching effectiveness, but also enables the development of digital literacies, which are essential in the modern academic environment.

In higher education, these chatbots answer students' questions, provide feedback and assessments, and facilitate communication and collaboration among students and educators. Technologies such as dialog agents, virtual assistants, automated chatbots, and conversational agents have become an integral part of the educational process, supporting both students and educators in their daily tasks [5]. For example, according to Hee-Seung Henry Bom [6], ChatGPT can support students by generating texts, answering questions, and helping them develop essay concepts and structures. Research suggests that the use of this tool helps to reduce the time needed for information retrieval and supports the development of writing and analytical skills. As a result, ChatGPT can serve as an interactive learning assistant, allowing students to access knowledge faster and learn more effectively.

Despite the numerous benefits associated with the use of artificial intelligence tools in higher education, their integration raises several ethical concerns. Tools such as ChatGPT, Co-Pilot, Gemini, and Jasper AI, while enabling the generation of high-quality essays and other academic content without the intellectual engagement required in traditional writing, can lead to misuse [7]. Students may opt for ready-made solutions instead of engaging in the independent creative and analytical process, which can threaten academic integrity and originality. Furthermore, there is a risk that these algorithms may

generate biased content, which could hinder the critical evaluation of presented information and lead to the formation of biases. As noted by Rane and colleagues [8], over-reliance on AI technology can limit students' ability to solve problems independently and develop analytical skills.

On the other hand, numerous studies suggest that chatbots, when properly implemented and regulated, can significantly contribute to improving the quality of education. Fauzi and colleagues [9] observe that technologies like ChatGPT can support the teaching process by helping generate examples, develop argumentative skills, and assist in text editing. Additionally, they can help reduce educational barriers by providing academic resources to students from marginalised groups or those with limited access to traditional forms of educational support.

However, it is important to note that the use of AI tools comes with the risk of bypassing traditional plagiarism detection systems, which presents a challenge to ensure the originality of students' intellectual creations. There are also technologies that refine and embellish content created by AI, further complicating the detection of dishonest use of these tools [10]. From the student perspective, while the speed of access to information and assistance in content creation improves learning efficiency, concerns arise about the decline in the quality of independently produced work and the risk of over-reliance on technology, which could limit the development of critical thinking [11]. The decision to use these tools often depends on clear guidelines from educational institutions and transparency regarding policies on the ethical use of AI technology.

3 Current Study

Interdisciplinary research on the development of artificial intelligence and LLMs-based tools is becoming an integral part of educational research. On the one hand, researchers are trying to explore the applicability of AI-based tools in education [12] while on the other hand, they are observing how AI is shaping education [13], taking into account the multitude of ethical issues that arise in this area [7]. As AI continues to develop, it is therefore necessary to carry out ongoing research to enable its informed integration into the educational process. Secondly, it is important to monitor and document the development of AI in an educational context, so that future generations of researchers can analyse the emerging stages of this process. In order to fill the data gap in this area, our study has a twofold character. It focuses on a specific sample of pedagogy students (learners and, at the same time, future teachers) and aims to understand how AI tools could influence future teaching practices. In line with the stated aim, our study is guided by the following research questions:

1. What are the patterns of use of AI-based tools by students at a pedagogical university?
2. How do students rate their knowledge and willingness to learn about the use of chatbots and other AI-language tools?
3. What are students' attitudes towards using AI-based tools?
4. What concerns do students have about using AI-based tools?
5. What are students' perceptions of the ethical issues that arise from the use of AI tools?

4 Method

4.1 Procedure

There are 3'832 students at the university we affiliate. Among them, we distributed a survey questionnaire in March 2025, which was made available via an anonymous link or a quick response (QR) code to be used according to the respondent's preference. The students were informed that the survey was voluntary and anonymous, and that they could leave at any time without facing any consequences. The questionnaire contained five blocks of closed questions and five open-ended questions. Two blocks of closed questions addressed the familiarity and frequency of use of selected AI chatbots and the familiarity and frequency of use of tools based on large language models [14], with a four-point scale ranging from 1 (unfamiliar) to 4 (familiar and frequently used). One block of closed-ended questions addressed the frequency of use of selected areas of AI chatbot functionality, with the 18 author-developed items rated on a five-point scale ranging from 1 (never use) to 5 (use regularly). Additional two blocks of closed questions addressed the knowledge, attitude, concern, use of chatbots/tools based on large language models, perceived ethics items rated on a five-point scale ranging from 1 (strongly disagree) to 5 (strongly agree) [14, 15]. Five open-ended questions developed by the authors covered topics such as experiences with tools built on LLMs, opinions on the integration of artificial intelligence in formal education, challenges and dilemmas of using AI in higher education.

4.2 Participants

There were 445 participants who took part in the study. In the first step, we discarded records with less than 50% completion. The final stage of the analysis incorporated data collected from 408 participants ($M_{\text{age}} = 22.44$, $SD = 4.63$, range 18–51), representing 10.6% of those eligible. The sample included 373 women (91.4%, $M_{\text{age}} = 22.5$, $SD = 4.7$, range 18–51), 28 men (6.9%, $M_{\text{age}} = 22.7$, $SD = 4.1$, range 19–35) and 7 (1.7%) of those who preferred not to provide gender information. The predominance of women is due to the structure of students body that is typical of pedagogical studies. Regarding the field of study, the sample consisted of 179 students of pedagogy with a teaching qualification (43.9%), 144 students of pedagogy without a teaching qualification (35.3%), 81 students of psychology (19.9%) and 4 other students (philosophy, sociology). Regarding the level of study, there were 116 students (28.4%) of bachelor's level (1st cycle), 269 students (65.9%) of master's level (long cycle) and 23 students (5.6%) of master's level (2nd cycle). Regarding the study mode, there were 287 full-time students (70.3%, studying during week days) and 121 part-time students (29.7%, studying during weekends).

4.3 Data Analysis Strategy

The present study adopts an exploratory analysis framework [16] within a mixed methods approach. The methodology employed incorporates both closed (quantitative) and open (qualitative) survey responses, with the objective of providing a more comprehensive understanding of the research problem [17]. The research material used for the analyses

was obtained from a questionnaire drawn up and completed in Polish. The quotations employed to illustrate the findings of the analyses were subsequently translated into English.

We used PS IMAGO PRO 10.0 with the IBM SPSS Statistics 29.0.2 analytical engine to analyse the quantitative data. To assess the differences between groups (pedagogy with teaching qualifications students, pedagogy without teaching qualifications students, students of social sciences but others than pedagogy) in using chatbots we employed non-parametric Kruskal-Wallis test and Mann-Whitney U-tests.

In order to explore the depth of context and enhance credibility, and thereby uncover students' underlying reasons, motivations and experiences of using AI-based tools, an analysis was conducted of the answers to open-ended questions. Emerging themes have been examined by individual researchers (the authors of the article) and representative quotes have been selected to illustrate key findings. The employment of qualitative analysis in conjunction with quantitative analysis has been demonstrated to counteract the respective limitations of each individual method, thus engendering a more comprehensive analysis [17].

5 Results

Participants of the study indicated that they know and use chatbots built on LLMs, however ChatGPT is the tool with which they have the most familiarity (50.5% of participants know and use it frequently, 38.0% know but use it rarely, 1.2% do not know it, see Fig. 1). They were more familiar with other AI language tools such as language translation tools, for example Google Translate, DeepL Translator (63.4% of students know and use them frequently, 28.5% know but use them rarely, 2.9% do not know such tools), language learning applications, for example Duolingo, Babbel (21.4% of students know and use them frequently, 34.2% know but use them rarely, 3.9% do not know such applications) or speech-to-text transcription tools, for example automatic YouTube subtitling, TurboScribe, Note, TranscribeMe (21.4% of students know and use them frequently, 34.2% know but use them rarely, 3.9% do not know such tools). To a lesser degree, students are familiar with and use online support tools to improve their writing, for example Grammarly, DeepL Write, LanguageTool (7.3% of students know and use them frequently, 13.8% know but use them rarely, 48.0% do not know such tools).

We were also interested in what purposes and areas students use chatbots for. The educational issues were the most popular. Among these, the most common (64.6%) was clarifying complex topics from classes, lectures, read articles (14.5% of students use it regularly, 24.3% often, 25.8% sometimes). Almost half of the students create summaries of long reading material such as articles, book chapters, handbooks (10.8% of students use this feature regularly, 16.0% use it often, 21.4% use it from time to time). 43.5% of students declare that they generate ready answers during classes, homework (8.1% regularly, 13.5% often, 21.9% from time to time). Every third student uses chatbots to make proofreading of self-written texts (7.4% regularly, 15.2% often, 16.0% from time to time), to make language translation (7.4% regularly, 10.3% often, 17.0% from time to time) and to generate content when writing essays (5.4% regularly, 8.6% often,

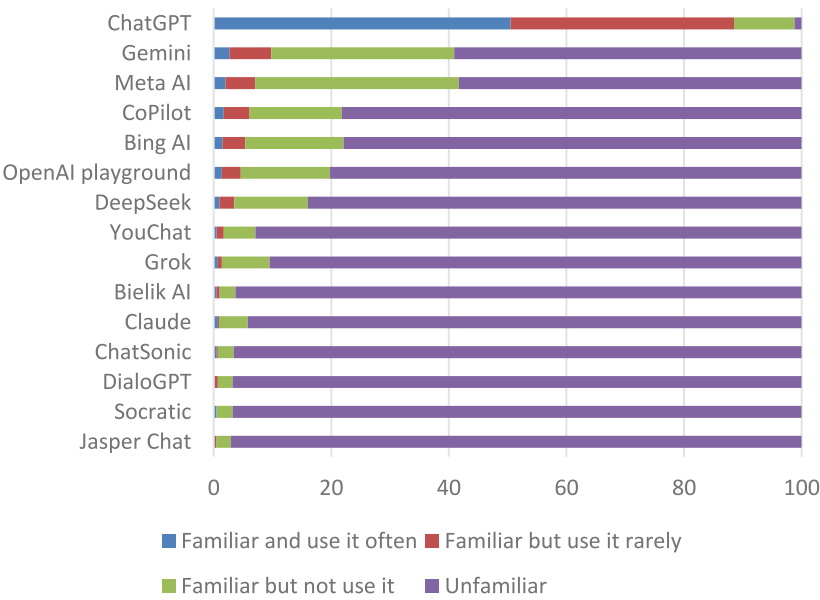


Fig. 1. Familiarity and usage of AI chatbots among students (in %). Missing answers were ignored, $N = 408$.

18.9% from time to time). A Kruskal-Wallis test indicated that there was only one slight significant difference in area of using chatbots between students representing various study fields, $H(2) = 7.68, p = .022$. Subsequent Mann-Whitney U tests indicated that students of not pedagogical field of study (psychology and others) use chatbots to generating of ready-made answers during examinations ($M = 1.42, SD = 0.92$) rarely than students of pedagogy with teaching qualifications ($M = 1.63, SD = 0.96, U = 6488.5, p = .022, r = 0.14$) and students of pedagogy without teaching qualifications ($M = 1.80, SD = 1.16, U = 4984.0, p = .007, r = 0.18$). Other differences in chatbot usage domains were not significant, $p > .134$.

In addition to the closed questions, students could add an extra answer to an open-ended question, indicating other purposes for which they use chatbots. They pointed out that they “use the following mainly in the context of support for their studies”, “correcting punctuation due to dyslexia”, “generating a plan for writing written work”, “searching for synonyms that match the text”, “learning – preparation of tests to check knowledge of the particular material” “assistance in giving examples of arguments”, “invitations, birthday wishes, emails to professors”, “finding sources for works”, “help with writing lesson plans, devising games”, “ideas for scout meetings scenarios”, “writing applications, official emails”, “generating content for Instagram posts”, “what to improve to better develop my passions, such as tips on painting techniques”, “to care for plants”, “when I have trouble setting something up on the computer, when something breaks down”, “for entertainment at most, I ask to write a poem”, “identifying songs”, “curiosity as to how it will answer various questions”, “life advice (you could say psychological) as mental support”.

We then explored responses to questions about students' stated knowledge and willingness to learn about chatbots other AI-language tools. More than half of the students claimed to know how chatbots work (7.7% strongly agree, 46.3% rather agree). A Kruskal-Wallis test showed that there was no significant difference between students representing different fields of study, $p = .656$. It is interesting to note that even though students claim to have a fairly good knowledge of chatbots, only a quarter of them have read any articles or research papers on chatbots (3.8% strongly agree, 22.8% rather agree), with no difference between students in terms of field of study, $p = .278$. At the same time, they believe that students should participate in training on the use of AI (27.7% strongly agree, 35.1% rather agree), with no difference between students, $p = .947$.

More than half of students have a positive attitude towards the use of chatbots in education (14.4% strongly agree, 40.8% rather agree), without difference between students, $p = .755$. Only one in five respondents believes that the use of chatbots should be banned in educational establishments (7.2% strongly agree, 11.5% rather agree), without difference between students, $p = .058$, however they are concerned about how chatbots will impact education in the future (39.7% strongly agree, 37.7% rather agree), $p = .574$. Respondents have the impression that the use of chatbots is widespread among their colleagues (49.5% strongly agree, 35.6% rather agree). A Kruskal-Wallis test indicated that there was significant difference between two groups of students, $H(2) = 8.15$, $p = .017$. Subsequent Mann-Whitney U tests indicated that use of chatbots is less widespread among students of not pedagogical field of study (psychology and others, $M = 4.11$, $SD = 0.83$) than students of pedagogy with teaching qualifications ($M = 4.39$, $SD = 0.83$, $U = 5833.5$, $p = .005$, $r = 0.17$). Few respondents are in agreement that their teachers or university have rules or guidelines for the responsible use of chatbots (3.8% strongly agree, 14.9% rather agree). This agreement is stronger in case of students of pedagogy without teaching qualifications ($M = 2.91$, $SD = 1.07$) rather than students of not pedagogical field of study ($M = 2.52$, $SD = 0.96$, $U = 4423.0$, $p = .008$, $r = 0.18$) and students of pedagogy with teaching qualifications ($M = 2.54$, $SD = 1.07$, $U = 9140.0$, $p = .002$, $r = 0.17$).

Looking at the ethical issues, more than 60% of those surveyed believe that using chatbots to complete assignments and exams is cheating (35.4% strongly agree, 32.1% rather agree). Almost half of the respondents feel privacy concerns when using a chatbot (15.4% strongly agree, 31.0% rather agree). Almost half of the respondents feel security concerns when using a chatbot (13.1% strongly agree, 27.9% rather agree). More than 70% of those surveyed believe that chatbots can be used to manipulate information or mislead students (30.0% strongly agree, 40.8% rather agree). Students also believe that chatbots could increase inequalities in education (21.0% strongly agree, 33.1% rather agree). A Kruskal-Wallis tests indicated that there were no significant differences between students representing various study fields, $p > .063$.

6 Discussion

The results of our study confirm that chatbots, particularly those based on large language models (LLMs) such as ChatGPT, have become common tools supporting students' academic activities. The vast majority of respondents reported using chatbots for educational purposes, which aligns with trends observed in the existing literature [5, 14, 18, 19]. Chatbots are most frequently used to simplify and explain complex content, highlighting their potential in supporting self-regulated learning and the development of cognitive skills. Importantly, open-ended responses revealed that although students appreciate chatbots as tools that support learning, they tend to approach chatbot-generated content with caution. As one respondent noted: "I mainly use it for low-stakes information, and I personally verify anything that requires deeper validation". This suggests a certain level of user awareness regarding the limitations of AI-generated outputs.

At the same time, the study identified ethically questionable practices, including the use of chatbots to generate ready-made answers during exams [20, 21]. Significant differences were observed between students from pedagogical and non-pedagogical fields: pedagogy students were more likely to admit to practices that may raise ethical concerns. This could be attributed to greater familiarity with AI tools or different perceptions of the role of technology in education. Nevertheless, it is worth noting that some students firmly rejected the use of chatbots for such purposes on moral grounds, as exemplified by one respondent who stated: "No, using chatbots goes against my moral principles". This highlights the diversity of students' attitudes toward the use of AI.

Another important finding is the identified need for training in the conscious and ethical use of AI in education. Although many students actively use chatbots, only a small portion have engaged with academic literature on the topic. This may suggest that students' knowledge of AI in education is often superficial and would benefit from further development.

On the one hand, respondents generally expressed positive views regarding the role of chatbots in improving access to information and supporting the learning process. On the other hand, they voiced concerns related to the influence of AI on critical thinking, data privacy, and educational equity. These concerns align with findings from other studies [15, 22] and emphasise the need for clear regulatory policies on the use of AI in education. Similarly, our results show that over half of the students expressed a positive attitude towards the use of chatbots in education (14.4% strongly agreed and 40.8% rather agreed). This tendency towards optimism is also reflected in studies conducted in Slovenia, where 89% of students reported a positive attitude towards the use of AI tools in education [19], and only a small percentage (2%) disagreed, while 9% remained neutral.

Regarding the perceived prevalence of chatbot use among peers, our respondents also indicated that this phenomenon is widespread (49.5% strongly agreed and 35.6% rather agreed). Similar observations were made in the Slovenian study, where 93% of students agreed that AI tools were commonly used by their classmates [19]. These results highlight the growing normalisation of AI technologies in academic environments. Taken together, the findings point to a general openness among students towards AI-driven tools in education, while also underscoring the importance of addressing critical challenges through research, policy development, and the implementation of best practices [23, 24].

Furthermore, the reported lack of clear guidelines and policies from universities and educators deserves particular attention. The absence of institutional directives may lead to inconsistent practices and reinforce ethical ambiguity among students. For instance, some students may regard the use of ChatGPT for paraphrasing texts as a legitimate form of independent academic work, whereas others, concerned about potentially breaching institutional regulations, may choose to forgo its use entirely, thereby missing opportunities to develop key competencies aligned with the demands of the contemporary labour market. Our findings align with those [25] who emphasise that higher education institutions should provide clearer policies and better educational programs on the responsible use of AI tools in academic work.

In the absence of official policies from educational institutions, educators may also adopt their own, often divergent, approaches to the use of AI, resulting in a lack of consistency in the assessment of students' skills. Consequently, students do not receive a clear message regarding the digital competencies expected of them, nor guidance on how to develop the ability to use AI tools critically, ethically, and creatively in both their academic pursuits and future professional careers. The absence of clear standards may also reinforce the illusion that AI tools pose solely a threat to the intellectual creation originality, rather than representing a potential means of supporting the development of competencies such as data analysis, problem-solving, and written communication. Yet, when appropriately introduced, AI can serve as an educational tool that fosters personalised learning, enhances metacognitive reflection, and improves cognitive efficiency. For this reason, the development and implementation of clear, transparent institutional guidelines for AI use not only help to structure educational practices but also enable the systematic cultivation of key future-oriented competencies, particularly in the areas of critical thinking, ethical technology use, and informed knowledge management.

7 Conclusion

This study explores the perspectives of Polish pedagogical university students on the use of chatbots in higher education, and provides key insights into their perceived benefits, challenges, and potential applications. The findings reveal that while chatbots are generally viewed as useful tools for enhancing learning efficiency, accessibility, and organisational support, concerns remain regarding the ethical issues associated with their use. The results indicate that students predominantly utilise ChatGPT for academic purposes, with limited application in daily life contexts. Notably, non-pedagogy students showed a greater tendency than their education-focused peers to employ the tool for generating pre-written exam responses. Although students self-reported a reasonable familiarity with chatbot technology, only a minority had engaged with scholarly literature or studies on the subject.

The research further examines ethical dimensions of chatbot usage in academic education. A small proportion of respondents perceived their institutions or educators as having established policies for responsible chatbot use, with non-pedagogy students and pedagogy students rating institutional preparedness in this area particularly low. Nearly half of participants raised concerns about privacy and data security when interacting with chatbots, while 70% expressed concern that chatbots could be exploited to disseminate manipulated or misleading information. Additionally, students identified potential risks of chatbots exacerbating educational disparities.

The analysis of the survey results presented here has some limitations. The survey was conducted in a selected public Pedagogical University in Warsaw, and therefore its results reflect the institutional context of the participants, which limits the possibility of generalisation to a wider student population. It is possible that replication of the study in a pedagogical university in another city or in a non-public university will show different trends, so it would be advisable to continue the study among students of pedagogical studies representing other academic units, which would allow generalisation of the results or show broader trends in the area of use of AI tools in higher education.

The data were collected at a single time point, preventing analysis of how chatbot usage or ethical perceptions evolve over time or in response to institutional interventions, than future research could include also that aspect. While ethical concerns were identified, the study did not delve deeply into contextual factors such as cultural norms, disciplinary differences, that might shape students' ethical decision-making. Another issue is that claims about students' familiarity with chatbots were based on self-assessment rather than objective measures, potentially overestimating actual understanding of AI technology. Another issue is the representativeness of the sample. Participation in the survey was voluntary and all students were encouraged to take part. However, some students opted out of the survey or withdrew during the survey, stating that they did not use chatbots. Therefore, the sample may not be fully reflective of those who are sceptical of artificial intelligence.

A notable implication of this study is the need for balanced integration, chatbots should be designed to complement, not compete with, human educators. Future research could explore adaptive chatbot functionalities that better simulate personalised tutoring while maintaining transparency about AI limitations. Additionally, institutional training on chatbot use may help maximise their educational benefits while mitigating scepticism. The uniqueness of the study lies in the selection of the research sample, such as the experiences, perceptions and attitudes towards AI-based tools of pedagogical students, those who will educate the next generation in the near future. The students' perspective, as direct users of the technology, provides a valuable reference point for shaping future strategies for the implementation of AI in higher education.

In conclusion, chatbots hold significant promise in higher education, but their successful adoption depends on thoughtful implementation that aligns with pedagogical goals and student expectations. As AI technology evolves, ongoing evaluation and stakeholder feedback will be essential to refine their role in education.

Authors' Contribution. Conceptualisation: E.D.; Methodology: E.D.; Formal analysis: E.D.; Investigation: E.D., E.M.-K., and J.W.-M.; Data curation: E.D.; Writing—original draft preparation: E.D., E.M.-K., and J.W.-M.; Review and editing: E.D., E.M.-K., and J.W.-M.; Supervision: E.D. All authors have read and agreed to the published version of the manuscript.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Yigci, D., Eryilmaz, M., Yetisen, A.K., Tasoglu, S., Ozcan, A.: Large language model-based chatbots in higher education. *Ad. Intell. Syst.* **6**(5), 2400429 (2024). <https://doi.org/10.1002/aisy.202400429>
2. Selwyn, N.: *Should Robots Replace Teachers? AI and the Future of Education*. John Wiley & Sons, Hoboken (2019)
3. Dempere, J., Modugu, K., Hesham, A., Ramasamy, L.K.: The impact of ChatGPT on higher education. *Front. Educ.* **8**, 1206936 (2023). <https://doi.org/10.3389/feduc.2023.1206936>
4. Silva, C.A.G.D., Ramos, F.N., de Moraes, R.V., Santos, E.L.D.: ChatGPT: challenges and benefits in software programming for higher education. *Sustainability* **16**(3), 1245 (2024). <https://doi.org/10.3390/su16031245>
5. Ilieva, G., Yankova, T., Klisarova-Belcheva, S., Dimitrov, A., Bratkov, M., Angelov, D.: Effects of generative chatbots in higher education. *Information* **14**(9), 492 (2023). <https://doi.org/10.3390/info14090492>
6. Bom, H.S.H.: Exploring the opportunities and challenges of ChatGPT in academic writing: a roundtable discussion. *Nucl. Med. Mol. Imaging* **57**(4), 165–167 (2023). <https://doi.org/10.1007/s13139-023-00809-2>
7. Oyedokun, G.E.: Artificial intelligence and academic integrity: a new frontier in combating emerging threats. In: Okebukola, P.A. (eds.) *A handbook on Artificial Intelligence and Quality Higher Education*, vol. 3. pp. 189–196. Sterling Publishers, New Delhi (2025)
8. Rane, N. L., Choudhary, S. P., Tawde, A., Rane, J.: ChatGPT is not capable of serving as an author: Ethical concerns and challenges of large language models in education. In: Okebukola, P.A. (ed.) *Proceedings of the 2023 International Conference on AI in Education*, pp. 851–874. Springer, Heidelberg (2023). <https://www.doi.org/https://doi.org/10.56726/IRJMETS45212>
9. Fauzi, F., Tuhuteru, L., Sampe, F., Ausat, A.M.A., Hatta, H.R.: Analysing the role of ChatGPT in improving student productivity in higher education. *J. Educ.* **5**(4), 14886–14891 (2023). <https://doi.org/10.31004/joe.v5i4.2563>
10. Dakakni, D., Safa, N.: Artificial intelligence in the L2 classroom: implications and challenges on ethics and equity in higher education: a 21st century Pandora's box. *Comput. Educ.: Artif. Intell.* **5**, 100179 (2023). <https://doi.org/10.1016/j.caeai.2023.100179>
11. Paseri, L., Durante, M.: Examining epistemological challenges of large language models in law. *Cambridge Forum on AI: Law Governance* **1**, e7 (2025). <https://doi.org/10.1017/cfl.2024.7>
12. Fuchs, K., Aguilos, V.: Integrating artificial intelligence in higher education: empirical insights from students about using ChatGPT. *Int. J. Inform. Educ. Technol.* **13**(9), 1365–1371 (2023). <https://doi.org/10.18178/ijiet.2023.13.9.1939>
13. Grassini, S.: Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings. *Educ. Sci.* **13**(7), 692 (2023). <https://doi.org/10.3390/educsci13070692>

14. Malmström, H., Stöhr, C., Ou, A.W.: Chatbots and other AI for learning: a survey of use and views among university students in Sweden. *Chalmers Stud. Commun. Learn. Higher Educ.* (2023). <https://doi.org/10.17196/cls.csclhe/2023/01>
15. Gonzalo Acosta-Enriquez, B., et al.: Knowledge, attitudes, and perceived ethics regarding the use of ChatGPT among Generation Z university students. *Int. J. Educ. Integr.* (2024). <https://doi.org/10.1007/s40979-024-00157-4>
16. Hochbein, C., Smeaton, K.S.: An exploratory analysis of the prevalence of quantitative research concepts in journal articles. *Int. J. Educ. Policy Leadership* (2018). <https://doi.org/10.22230/ijepl.2018v13n11a765>
17. Lund, T.: Combining qualitative and quantitative approaches: some arguments for mixed methods research. *Scand. J. Educ. Res.* **56**(2), 155–165 (2012). <https://doi.org/10.1080/00313831.2011.568674>
18. Moussa, N., Bentoumi, R., Saali, T.: Promoting student success with neutrosophic sets: artificial intelligence and student engagement in higher education context. *Int. J. Neutrosophic Sci.* **23**(1), 238–248 (2024). <https://doi.org/10.54216/IJNS.230121>
19. Fošner, A.: University students' attitudes and perceptions towards AI tools: implications for sustainable educational practices. *Sustainability* **16**(19), 8668 (2024). <https://doi.org/10.3390/su16198668>
20. Cordero, D.A., Jr.: Addressing academic dishonesty caused by artificial intelligence tools. *J. Public Health* **46**(2), e310–e310 (2024). <https://doi.org/10.1093/pubmed/fdad233>
21. Cotton, D.R., Cotton, P.A., Shipway, J.R.: Chatting and cheating: ensuring academic integrity in the era of ChatGPT. *Innov. Educ. Teach. Int.* **61**(2), 228–239 (2024). <https://doi.org/10.1080/14703297.2023.2190148>
22. Ogugua, D., Yoon, S.N., Lee, D.: Academic integrity in a digital era: Should the use of ChatGPT be banned in schools? *Global Bus. Financ. Rev.* **28**(7), 1–10 (2023)
23. Farhi, F., Jeljeli, R., Aburezeq, I., Dweikat, F.F., Al-shami, S.A., Slamene, R.: Analyzing the students' views, concerns, and perceived ethics about ChatGPT usage. *Comput. Educ.: Artific. Intell.* **5**, 100180 (2023)
24. Mendoza, S., Sánchez-Adame, L.M., Urquiza-Yllescas, J.F., González-Beltrán, B.A., Decouchant, D.: A model to develop chatbots for assisting the teaching and learning process. *Sensors* **22**(15), 5532 (2022). <https://doi.org/10.3390/s22155532>
25. Singh, H., Tayarani-Najaran, M.-H., Yaqoob, M.: Exploring computer science students' perception of ChatGPT in higher education: a descriptive and correlation study. *Educ. Sci.* **13**(9), 924 (2023). <https://doi.org/10.3390/educsci13090924>



Parental Memory and Digital Traces of School Closures During the COVID-19 Pandemic: What Is Remembered, What Fades, and What Is Left Behind

Akiko Orita^(✉) 

Kanto Gakuin University, Yokohama 231-8328, Japan
akiko.orita3@gmail.com

Abstract. This study examines how Japanese parents remember and record their experiences during the COVID-19 pandemic, particularly focusing on school closures in spring 2020. Using a mixed-methods approach—semi-structured interviews ($n = 10$) and an online survey ($n = 200$)—the research explores what is remembered, what fades, and what remains in digital form. Many participants misremembered the timing of school closures and the State of Emergency, often recalling them as starting earlier and ending later than in reality. Women were more likely to perceive earlier disruption, reflecting potential gendered burdens in caregiving. Preventive behaviors like handwashing persisted, but scientific understanding often faded. While 71% used social media, few actively documented their experiences. Notably, those who found the period emotionally painful were more inclined to want their experiences preserved. These findings highlight a discrepancy between lived experience and administrative facts and raise ethical questions about whose voices—and which memories—are digitally preserved for the future.

Keywords: COVID-19 · Memory Distortion · Digital Legacy

1 Introduction

The COVID-19 pandemic drastically disrupted global daily life and social systems. People were asked to stay home, and some countries imposed strict lockdowns. As society resumed normalcy with the advent of vaccines and treatments, an important question emerges: how are those times remembered and recorded?

Studies have shown that during periods of social restriction, people's subjective sense of time became distorted, influencing how events were remembered (Martinelli et al., 2020; Ogden, 2020). In the digital age, these experiences were recorded not in letters or diaries, but through SNS posts, messages, and emails. Like the photographs and diaries from the Spanish flu era, these digital traces offer insight into lived experience. However, whether they will be preserved or lost remains uncertain (Doyle and Brubaker, 2023).

This study investigates how such memories are recalled and how digital traces are treated. It focuses on Japanese parents who experienced two major disruptions during

the early phase of the pandemic: nationwide school closures and the declaration of a State of Emergency.

Using a mixed-methods approach, we address the following questions:

RQ1: How accurately are the durations of the State of Emergency and school closures remembered?

RQ2: Are infection prevention behaviors still practiced?

RQ3: Were personal experiences recorded on SNS or similar platforms?

RQ4: How do people perceive the long-term preservation of such records?

This paper first reviews related research, then presents findings from qualitative interviews and a quantitative survey, followed by a discussion on how memory and digital records may or may not be inherited. The study contributes to scholarship on collective memory, digital legacy, and public health communication.

The research forms part of a broader project examining pandemic-era memory. While this paper focuses on Group B—parents of elementary school children—the experiences of Group A, university students who began college during campus closures, will be analyzed separately.

2 Related Works

2.1 Distortions of Memory and Time Perception During the Pandemic

Several studies have shown that people's sense of time was distorted during COVID-19 lockdowns. In France and the UK, time was often perceived as passing more slowly, influenced by boredom, sadness, stress, and reduced social interaction (Martinelli et al., 2020; Ogden, 2020). Similar findings have been reported in Spain, Argentina, and Brazil (Brenlla et al., 2022; Cravo et al., 2022). Interestingly, a follow-up study in France (Droit-Volet et al., 2023) found that while time felt slow during lockdown, people later remembered the period as shorter. These studies highlight a consistent link between psychological distress and subjective time perception.

2.2 Preservation of Personal Memories and Digital Legacy in the Digital Society

The concept of digital legacy has received growing attention as a means of preserving personal experiences and emotions from the pandemic as historical materials. In contemporary society, personal records—once kept in diaries or on paper—are now mostly digital, including SNS posts, messages, and emails. These records face uncertain futures: whether they are preserved or lost depends not only on individual intent, but also on cultural expectations, national practices, and still-developing legal frameworks surrounding data access, post-mortem privacy, and platform governance (Doyle and Brubaker, 2023; Kasket, 2020). For example, users in the United States are more likely to favor data preservation, while many in Japan express a preference for deletion after death (Orita, 2023). Even when individuals wish to preserve their digital legacy, there is no guarantee that commercial platforms will maintain long-term accessibility.

3 The COVID-19 and Social Context in Japan, Spring 2020

This section outlines the situation in Japan during spring 2020, providing essential context for this study.

In February 2020, the cruise ship *Diamond Princess* docked in Yokohama with a large number of COVID-19 infections onboard. The incident drew intense media coverage and heightened public awareness of the virus. Later that month, the Japanese government abruptly requested the closure of all elementary, junior high, and high schools nationwide. This sudden decision disrupted education and forced many parents to rearrange their work to care for children at home. In Japan's traditional gender role structure, where mothers typically bear primary childcare responsibilities, this burden fell disproportionately on women.

April marks the beginning of both the academic and fiscal year in Japan. Amid rising infections, the government declared a State of Emergency. Unlike legally enforced lockdowns in some Western countries, Japan's approach relied on voluntary compliance. Citizens were urged to stay home, and businesses were asked to suspend operations without legal penalties—an approach that defined Japan's pandemic policy.

Daycare centers requested that children of non-essential workers stay home. Some municipalities even discouraged children from playing in parks. Restaurants closed or shifted to takeout, shopping malls shut down, and streets emptied. Remote work became widespread, and schools distributed printed materials or held online classes. Daily news broadcasts tracked regional infection numbers. Authorities promoted the “Three Cs” (Closed spaces, Crowded places, Close-contact settings) avoidance strategy, along with mask-wearing, hand sanitizing, social distancing, and temperature checks. Infected individuals and close contacts were instructed to isolate for 14 days. Social stigma against those infected was also common.

The State of Emergency was lifted in phases, starting with less-affected regions, while others remained under restrictions longer. Additional frameworks like the “Priority Preventive Measures” were later introduced to manage ongoing waves. Citizens continued to be urged to maintain preventive behaviors even after restrictions eased.

4 Methodology

This study adopts a mixed-methods approach that combines qualitative and quantitative research in order to investigate how Japanese parents remembered and recorded their experiences during the early phase of the COVID-19 pandemic. The qualitative component explores individuals' deeper experiences and values, while the quantitative component aims to visualize trends and offer a degree of generalizability.

Both the qualitative and quantitative studies were designed to address the following research questions:

RQ1: How were the durations of the State of Emergency and school closures remembered, and how did these memories differ from the actual timeline?

RQ2: Are infection prevention measures still being practiced?

RQ3: Did participants record their experiences on social media or other platforms during the pandemic?

RQ4: How do participants feel about their data being preserved into the future?

The qualitative and quantitative findings reported here reflect an initial phase of analysis. Further in-depth coding of the qualitative data using KH Coder, as well as detailed statistical examination of public health behavior-related variables in the survey data, will be presented in separate papers.

4.1 Qualitative Research

Participants and Procedure

The qualitative research consisted of semi-structured interviews conducted via Zoom from November to December 2024. Participants were ten parents (five men and five women) who had children in the lower grades of elementary school (grades 1–3) in April 2020. Participants were recruited from a research monitor panel maintained by the company Asmarq, and each interview lasted approximately 40 to 60 min.

Ethical Considerations

This study was approved by the Research Ethics Committee on Human Subjects at Kanto Gakuin University (Approval No. H2024–3-3). Before participating, all interviewees were given a full explanation of the study's purpose and data protection policy, and they signed informed consent forms.

Interviews were recorded and transcribed using the AI transcription service Notta, and then manually corrected by the researcher.

Each participant was assigned a unique identifier (e.g., BM-1, BF-3), and no personally identifiable information was recorded or used.

4.2 Quantitative Research

Survey Design and Participants

The quantitative component was conducted via an online panel provided by Freeasy. A screening survey was administered on January 31, 2025, followed by the main survey on February 10, 2025. From 3,000 initial respondents, 200 parents who had children in elementary school during spring 2020 were selected. To ensure gender balance, the sample included 100 men and 100 women.

The survey was anonymous, and only participants who gave informed consent after being briefed on the study's purpose and data privacy policy were included. According to institutional guidelines, ethical approval was not required for anonymous online surveys.

Survey Content

The questionnaire mirrored the structure of the qualitative interviews and covered the following topics:

- Residence in spring 2020 (to determine applicability of the State of Emergency)
- Memory of the start and end of the State of Emergency and school closures
- Emotional impressions from that period (e.g., difficulty, confusion)
- Use of social media and digital platforms

Analytical Methods

The collected data were analyzed primarily using SPSS Statistics Ver.29, which was used for descriptive statistics, cross-tabulations, and chi-square tests. ChatGPT-4o was employed as an auxiliary tool to support the interpretation of complex tabular data and to generate exploratory summaries, under the authors' supervision and judgment.

5 Results

5.1 Qualitative Results

This section presents the findings from semi-structured interviews with 10 participants (5 male and 5 female), organized according to the four research questions (RQ1–RQ4). The participant identifiers follow the format BM1–BM5 for male participants and BF1–BF5 for female participants. Illustrative quotes are used to highlight key patterns that emerged in their narratives.

1) Discrepancies Between Actual and Remembered Periods of the State of Emergency and School Closures.

Although the State of Emergency in Japan was in effect between April 7 and May 25, 2020, most participants remembered it as starting earlier than it actually did. Despite being informed at the beginning of the interview that the declaration was issued in April, many referred to March—likely due to the prior nationwide school closures—and reported a general sense of restricted mobility from that time.

“I’m not sure if the entrance ceremony was in April or May.” (BM-2).

“I feel like it was in March or even February.” (BF-3).

Regarding the end of the State of Emergency and school closures, participants also tended to recall a timeline later than the actual dates. Some even cited the autumn of 2020 or the following year.

“I don’t really remember. It felt like it ended about a year later.” (BM-3).

“I don’t remember at all. Maybe around September? I feel like there was a cycle of declarations and lifting of restrictions, so it didn’t really end all at once.” (BF-5).

“Was it May 2023 when it changed from Category II to Category V under the Infectious Disease Act? I might be mixing things up.” (BM-5).

These responses suggest that participants tended to remember the period of restrictions as longer than it actually was, reflecting a subjective distortion of time.

2) Continuation and Changes in Infection Prevention Measures.

While some participants maintained certain COVID-19 prevention behaviors, many had abandoned most of them. Only 2 of the 10 interviewees reported continuing to wear masks—once a near-universal behavior during the pandemic.

“Masks didn’t feel that meaningful in terms of prevention. It was more like social pressure—everyone else was wearing one, so I felt I had to. I don’t wear one at all anymore. But handwashing has become a habit.” (BM-2).

“What I’ve stopped is wearing a mask. What I’ve continued is alcohol sanitizing and handwashing.” (BF-3).

“I still sanitize my hands and wear a mask, but I don’t really avoid the ‘Three Cs’ anymore. I didn’t wear a mask in summer—it was just too hot.” (BM-3).

Although some behaviors such as handwashing persisted, participants often acknowledged that mask-wearing had been driven more by social conformity than by a clear understanding of its medical purpose. Notably, these interviews were conducted during a large influenza outbreak in Japan, yet the participants reported not wearing masks even in that context.

3) Social Media Use and Awareness of Digital Records.

Only one participant reported posting on Instagram during the pandemic. Others used social media primarily for information-gathering. Light interactions such as “liking” posts or communicating via LINE were observed, but no one else actively documented their own experiences.

“I just posted about things like baking sweets at home on Instagram.” (BF-4).

“I only looked at Twitter, maybe gave some likes. Mostly just followed trends.” (BF-5).

“I wondered how reliable TV broadcasts were, so I started checking what people were saying on YouTube and Twitter.” (BM-2).

In summary, none of the participants had preserved their personal pandemic experiences online, suggesting that their perspectives may be missing from the digital historical record.

3) Attitudes Toward the Future Preservation of Personal Data.

The only participant who had posted to Instagram expressed discomfort about having that content preserved:

“I’d rather it be deleted. It’s embarrassing that it’s still there.” (BF-4).

Others, despite not posting publicly, generally expressed openness to having their experiences preserved for future generations—especially in the context of historical documentation.

“I don’t care about leaving something about myself, but I think it would be good if it’s written that it was a time when families spent more time together, even though it was hard.” (BM-5).

“If it’s 100 years from now, sure—please read it. I’d want people to know what we thought and did back then, both the good and the bad.” (BF-5).

“As long as it doesn’t cause trouble for my children. I wouldn’t want them to be told things like ‘your dad did this.’” (BM-2).

Interestingly, those who had not posted on social media were more open to the idea of preservation than the one participant who had. However, many expressed concern about how their data might affect their children in the future, indicating that intergenerational implications influenced their attitudes toward data permanence.

5.2 Quantitative Results (Descriptive Statistics)

This section presents the results of the online survey conducted in February 2025, based on 200 valid responses (100 male and 100 female participants). The respondents were parents living with elementary school children during the spring of 2020. The survey content aligned closely with the qualitative interview and focused on memory, behavior, documentation, and attitudes toward record preservation.

Memory of the State of Emergency and School Closures

Respondents were asked when the State of Emergency and school closures began and ended in their region. Responses were compared with actual dates of government announcements.

In Japan, the State of Emergency was first declared in seven prefectures on April 7, 2020, and expanded nationwide on April 16. It was lifted in phases: May 14 (39 prefectures), May 21 (3 Kansai prefectures), and May 25 (remaining areas). The start and end of school closures varied by region. Responses were categorized as “accurate,” “earlier than actual,” or “later than actual.” The results are summarized in Tables 1 and 2:

Table 1. Perceptions for Start and End of State of Emergency

	Start of SoE	End of SoE
Earlier than actual	61.0%	35.5%
Accurate	12.0%	24.5%
Later than actual	27.0%	40.0%

For the State of Emergency, 61.0% of respondents recalled it beginning earlier than it actually did, and 40.0% recalled it ending later. The most common memory for the start was “late March” (45.0%). For the end, the most frequent answer was “after August” (25.0%). These discrepancies suggest that respondents remembered the subjective experience of restriction more strongly than the official announcements. Perceived start dates did not significantly differ across regions, implying that memory was influenced more by social atmosphere than by regional policy differences. However, gender showed a significant difference: Pearson’s $\chi^2 = 12.308$, $df = 4$, $p = .015$. Women (57%) were more likely than men (33%) to remember the start as “late March,” possibly reflecting the disproportionate burden on mothers caused by sudden school closures.

Table 2. Perceptions for Start and End of School Closure

	Start of Closure	End of Closure
Earlier than actual	36.5%	11.5%
Accurate	30.0%	11.5%
Later than actual	33.5%	77.0%
Unknown	0.5%	0.5%

Regarding school closures, 33.5% remembered them starting earlier than actual, while 77.0% believed they ended later. A significant gender difference was again observed in the memory of the start: Pearson’s $\chi^2 = 31.829$, $df = 5$, $p < .001$; women 51.0%, men 16.0%. However, for the end of school closures, gender differences were not statistically significant.

The perceived length of these periods is summarized in Fig. 1:

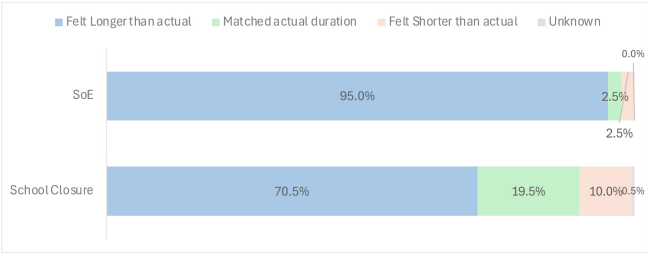


Fig. 1. Perceived length of these periods (N = 200)

2) Continued Infection Control Measures (Multiple Responses Allowed).

Figure 2 shows infection control measures that more than 50% of respondents reported practicing in 2020. Among these, handwashing with soap was the most consistently maintained behavior, with 60.0% of respondents still practicing it in 2024, almost unchanged from 59.5% in 2020. Mask-wearing also showed relatively high continuity, dropping from 69.5% in 2020 to 52.5% in 2024, indicating that more than half of respondents still wore masks. In contrast, behaviors such as avoiding the “Three Cs” (closed spaces, crowded places, and close-contact settings) and maintaining social distance, which had been practiced by 60.0% and 53.0% of respondents respectively in 2020, saw sharp declines in 2024 (21.5% and 22.0%, respectively).

This suggests that spatial and social distancing practices were largely discontinued, while personal hygiene practices such as handwashing remained ingrained.

Use of Social Media and Emotional State in 2020.

In the spring of 2020, 71% of respondents reported using social media. The most common purpose was information gathering (55.6%), followed by sharing others’ information

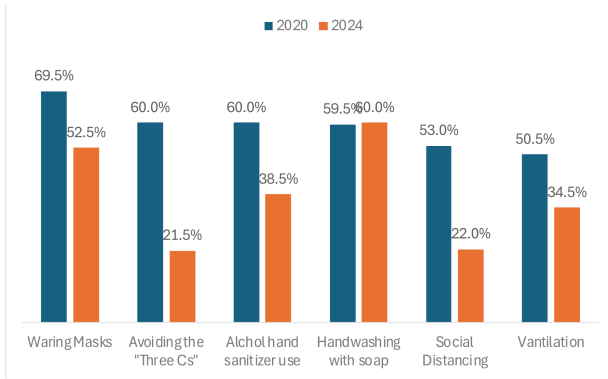


Fig. 2. Infection Control Measures in 2020 and 2024 (MA)

through actions such as retweets (41.5%), active posting (35.2%), and sharing personal updates within their social circle (31.6%). All purposes except information gathering showed statistically significant gender differences ($p < .01$), with women being more likely to post or share content, while men were more likely not to use social media at all. The specific types of posts shared are summarized in the Table 3.

Table 3. Topic of Social Media Posts (MA)

Content	
Information about the number of COVID-19 cases	16.2%
Information about infection prevention measures	19.0%
Information about the State of Emergency	15.5%
Information about school closures	14.8%
Information about children	15.5%
Information about work	18.3%
How to spend time at home ("stay-home" time)	15.5%
Personal stress, anxiety, or frustrations	16.2%
Encouragement for people in similar situations	14.1%
Entertainment such as movies, videos, or games	14.8%
Others	0.0%
None of Above	37.3%

Regarding emotional experiences during the pandemic, 35.5% of participants said it was "painful," 38.5% reported that it was "painful but included some positive aspects," 18.0% said it was "positive but with some difficulties," and 8.0% described it as simply "positive." No significant gender differences were found in emotional evaluation. However, cross-tabulations revealed that those who felt the period was painful were

significantly more likely to have engaged in both posting and sharing information on social media. Specifically, 31.5% of those who found it painful had posted (compared to 5.7% of those who found it positive), and 34.4% had shared information (compared to 15.3% of those who found it positive), both with p-values less than .01.

Attitudes Toward Future Preservation of Posts.

When asked whether they wanted their own or others' social media content preserved, 46.5% hoped their own posts would remain 10 years later, and 48.5% expressed the same for others' posts. For 100 years into the future, 38.5% supported the preservation of their own posts, while 47.0% supported the preservation of others' posts.

A statistically significant relationship was observed between emotional experience and willingness to preserve data. Those who had experienced emotional hardship were more likely to want both their own and others' content preserved, indicating a link between emotional intensity and a desire to record personal and collective history. Chi-square results were as follows:

Own posts preserved after 10 years (Q15-m1): $\chi^2 = 26.792$, $df = 9$, $p = .002$

Others' posts preserved after 10 years (Q15-m3): $\chi^2 = 30.637$, $df = 9$, $p < .001$

Others' posts preserved after 100 years (Q15-m6): $\chi^2 = 30.879$, $df = 9$, $p < .001$

These results suggest that those with more painful memories are significantly more motivated to preserve both personal and collective records. This tendency was also observed in the qualitative interviews.

6 Discussion

This section discusses the findings from both the qualitative and quantitative phases of the study, structured around the four research questions (RQs).

6.1 RQ1: How Were the Durations of the State of Emergency and School Closures Remembered, and How Did These Memories Differ from Actual Timelines?

Semi-structured interviews showed that many participants recalled the start of the State of Emergency earlier than its actual date of April 7—some placing it as early as February or March. The end was similarly misremembered, with some believing it lasted until September or even a full year later. These gaps suggest that memories were shaped more by the emotional climate and lifestyle disruptions than by official dates. For school closures, the use of phased or split attendance further blurred recall of when schools truly reopened.

Quantitative findings echoed these trends. Most respondents remembered both the State of Emergency and school closures as longer than they were. Start dates were frequently recalled as earlier than actual, with women significantly more likely to do so. Memory of the end was even more distorted—77% believed school closures ended later than they did.

These results underscore that public memories during crises are not anchored in official timelines but are instead influenced by subjective experience. This has implications for designing memory preservation and historical reconstruction methods that account for such distortions.

6.2 RQ2: Are Infection Prevention Measures Still Being Practiced? (from the Perspective of Public Health Knowledge Transmission)

The interview results showed that while some practices—such as handwashing and alcohol sanitization—had been maintained, behaviors like mask-wearing and social distancing had largely been discontinued. Notably, many participants described mask use during the pandemic as driven more by social pressure than scientific understanding. Even during the interviews—conducted amid a seasonal influenza outbreak—most reported not wearing masks, suggesting that public health knowledge had not been internalized.

The survey data supported these observations: while handwashing (59.5%) and mask-wearing (52.5%) remained relatively common in 2024, only 21.5% avoided the “Three Cs” and 22.0% maintained physical distancing. This decline implies that although hygienic habits persisted, spatial distancing practices had faded as everyday life returned to normal.

These findings underscore the need to consider not only what infection control measures were practiced, but also whether the public internalized the scientific rationale behind them. Without recording the reasons for actions—not just the actions themselves—there is a risk that future generations may fail to apply such knowledge in future crises. Archival strategies should therefore include not only behavioral records but also public health explanations to support intergenerational knowledge transmission.

6.3 RQ3: Did Participants Record Their Experiences on Social Media or Other Platforms During the Pandemic?

Only one interview participant reported actively posting to social media during the pandemic, and even then, the content was limited to light updates like homemade desserts. Others used SNS passively—for browsing or casual interactions via LINE or Twitter. Emotionally intense memories shared during interviews were not reflected in online posts, indicating that these experiences were not publicly documented.

In contrast, the quantitative survey revealed a higher proportion of social media engagement. Women were significantly more likely to post, share, or report their experiences, while men were more likely to abstain altogether. Furthermore, those who described the period as painful were significantly more likely to have posted or shared ($p < .01$). This suggests that emotional distress may motivate individuals to engage in digital expression as a form of coping or memory preservation.

These results highlight a major limitation of digital archives: while social media may appear to offer comprehensive historical records, it in fact reflects selective and unequal documentation. Emotional state, gender, and other factors shape who posts and whose voices are preserved. Archival and ethical practices should account for these gaps and ensure more inclusive digital memory frameworks.

6.4 RQ4: How Do Participants Feel About Their Data Being Preserved into the Future?

Interview data revealed mixed attitudes toward preserving one's digital records. While one participant said they would be embarrassed for their social media posts to remain accessible, others expressed willingness to preserve such data if it were anonymized or viewed as a historical record 100 years from now. Several noted that both the challenges and the positive aspects (e.g., increased family time, remote work) of the pandemic period should be conveyed.

The survey showed a stronger preference for preserving others' posts than one's own, especially for 100 years into the future. Moreover, individuals who described the experience as "painful" were significantly more likely to want their records to be preserved ($p < .01$). This suggests that the emotional weight of an experience may correlate with the desire to archive it. These findings imply that motivations for preservation are less about self-expression and more about ensuring that certain collective experiences—especially difficult ones—are not forgotten. The desire to record may stem from the need to bear witness, and thus archiving systems should accommodate both anonymity and emotional complexity. Future research and policy should address how to responsibly preserve digital records in ways that respect privacy, avoid undue burden on individuals, and acknowledge the uneven nature of who records and is recorded.

7 Conclusion

This study investigated how parents of elementary school children in Japan remembered and recorded their experiences during the early phase of the COVID-19 pandemic. Using a mixed-methods approach combining qualitative interviews and a nationwide quantitative survey, we addressed four key questions: the accuracy of memory, the continuation of preventive behaviors, the extent of social media sharing, and attitudes toward long-term preservation of records.

The central finding of this study is that memories of the pandemic were influenced more by everyday experiences and emotional impressions than by administrative facts. In particular, women tended to recall earlier start dates and longer durations, which is likely related to the tendency in Japanese society for caregiving and domestic responsibilities to fall more on women.

The study also revealed a gap between lived experience and digital recordkeeping. While social media is assumed to preserve public memory, many did not share their experiences. Those who found the period difficult were more willing to preserve records.

Positive and quiet moments—such as spending time with family—were rarely posted, while distressing experiences were more likely to be recorded. This raises ethical concerns about whose experiences are remembered and in what form, as digital archives may become skewed toward trauma.

This study has several limitations. The qualitative sample was small ($n = 10$) and drawn from an online panel, limiting demographic diversity. Both studies relied on retrospective self-reporting, which may increase the risk of memory distortion. Furthermore, the study focused exclusively on Japanese parents of school-aged children, and caution

should be taken when generalizing findings to other populations. Variations in digital literacy and individual personality traits may also have influenced the results.

Finally, while this study highlighted the potential fragility of public health memory, it did not propose concrete mechanisms for preserving and utilizing such knowledge in the future. Future interdisciplinary research is needed to explore how both personal and institutional memory systems can retain and transmit such lessons. In addition, digital platforms must develop tools that enable users to preserve their experiences in a privacy-conscious manner. Recognizing that memory is unevenly formed and that digital records are inherently selective, it is essential to promote more intentional and inclusive approaches to documenting diverse experiences.

Acknowledgments. This research is supported by JSPS Grant-in-Aid for Scientific Research (C) #22K12724.

References

- Brenlla, M.E., Carrera, M.V., Jáuregui, P.: Time experience and emotional state during the COVID-19 lockdown in Argentina. *Front. Psychol.* **13**, 876328 (2022). <https://doi.org/10.3389/fpsyg.2022.876328>
- Cravo, A.M., Silva, C.F., de Souza, L., de Freitas, L., Zilio, D., Walsh, V.: Time experience during social distancing: A longitudinal study during the first months of COVID-19 pandemic in Brazil. *Sci. Rep.* **12**, 14434 (2022). <https://doi.org/10.1038/s41598-022-18724-3>
- Doyle, D.T., Brubaker, J.R.: Digital Legacy: A Systematic Literature Review. *Proc. ACM Hum.-Comput. Interact.* **7**(CSCW2), 1–26 (2023). <https://doi.org/10.1145/3610103>
- Droit-Volet, S., Berthon, M., Ramos, D., Izaute, M., Dambrun, M., Gil, S.: Time and memory in COVID-19 lockdown: Time free, memory of what? *Timing & Time Perception* **11**(1), 70–93 (2023). <https://doi.org/10.1163/22134468-bja10065>
- Kasket, E.: Social media and digital afterlife. In: Savin-Baden, M., Mason-Robbie, V. (eds.) *Digital Afterlife*, pp. 21–34. CRC Press (2020)
- Martinelli, N., et al.: Time and emotion during lockdown and the Covid-19 epidemic: determinants of our experience of time. *Front. Psychol.* **11**, 616169 (2020)
- Ogden, R.: The passage of time during the UK COVID-19 lockdown. *PLoS ONE* **15**(7), e0235871 (2020). <https://doi.org/10.1371/journal.pone.0235871>
- Orita, A.: Retain or Delete? Intentions for Social Network Accounts After Death. In: Schott, G.R. (ed.) *The Art of Dying – 21st Century Depictions of Death and Dying*, pp. 89–114. Palgrave Macmillan, Cham (2023)



“There Are so Many” - Harms of Smart Homes

Sally Bagheri^(✉) 

Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden
sally.bagheri@mau.se

Abstract. In this paper, harms facilitated by smart homes are considered. The study aimed to provide an overview of the research area by contributing a map of smart home-facilitated harm and supporting future research. A systematic mapping study, including 48 published and peer-reviewed articles, identified eight preliminary categories of harm. These results were then presented to a focus group of nine smart home users to brainstorm additional categories, resulting in a comprehensive final map of several smart home-facilitated harms. The categories of harm identified in the study were wide-ranging, including domestic abuse, surveillance and privacy concerns, security breaches, algorithmic bias such as digital ageism, democratic risks, environmental degradation, and market monopolization by big tech companies. Suggestions for future work to address the societal and ethical implications of IoT life conclude the paper, mainly regarding smart homes but also their connection with smart cities.

Keywords: Smart Homes · Harms · Domestic Surveillance · Ethics.

1 Introduction

Smart homes are domestic spaces where people live complex and idiosyncratic lives amid the Internet of Things (IoT) [1]. They are a heterogeneous system of systems, deploy a wide range of IoT devices such as smart lights, smart doorbells, smart security cameras, and smart TVs, and interact with a diverse set of users, such as homeowners, tenants, children, neighbors, visitors, and domestic workers. In a sense, the smart home is an assemblage that includes both users and devices as nodes of agency (some IoT devices are meant to act autonomously based on predictive analyses as part of their functionality) and intersects with technology, policy, and practice [2]. Reports of unethical and, at times, dangerous uses of these systems are increasing, with two examples being an increased level of intrusive surveillance [3] and their facilitation of domestic abuse [4]. The literature on harms associated with smart homes is not as widespread as examining their benefits and optimization, although it is gaining increasing attention. Previous reviews have focused on the abovementioned dangers to tease out their nuances (for example, see [4]). Alternatively, scoping reviews have been conducted, including risks, vulnerabilities, and users’ agency in the discussion [5]. Such results tend to be too narrow (focusing on a specific harm) or too general (analyzing the types of publications and research disciplines).

Additionally, previous research reporting on smart home experiences has rarely considered harms of smart homes. Device-specific harms have been considered [6], as well as infrastructural challenges such as the wide array of security vulnerabilities smart home systems pose; some examples include denial of service attacks, eavesdropping, and vulnerable entry points, such as home routers, allowing access to entire smart home systems [7]. From a user perspective, ethical concerns related to the perception of smartness [8], survey studies based on negative experiences [9], and how smart homes might control and discipline users [10] hone in on aspects of smart home facilitated harm, but lack an overview related to the subject matter. Although smart homes potentially pose an array of harms, for example, privacy is commonly considered a significant concern related to smart homes [11], adoption rates are not ceasing; the penetration rate of smart homes is estimated to increase by 16.8% annually for the next couple of years¹. To make robust ethical evaluations, it is crucial to approach smart homes as assemblages of users, devices, and society at large and to consider harms they facilitate. Therefore, the study documented in this paper aims to provide an overview of the general opinions of smart home-facilitated harm to support future research efforts. The following research question (RQ) is posed: *What kinds of harm can smart homes facilitate?* We consider facilitating harm to be “acts of harm that are brought about when technical artefacts, through making a harmful end easier or possible to achieve, invite actors holding this end to actualize it using the artefact”[12]. In this sense, we are interested in identifying the forms of harm made easier or possible by smart homes.

2 Method

In this paper, a mixed-methods approach is used, shown in Fig. 1. The approach is selected to engage with two different kinds of data sources, previous research and user experience, to develop a comprehensive overview of the area.

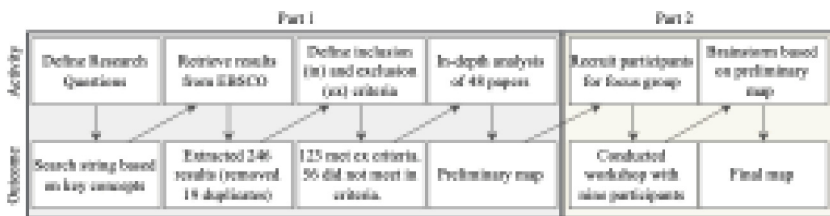


Fig. 1. Visual representation of the two-part research process: part one, a systematic mapping study [13], and part two, a mix of brainstorming and focus groups with end-users [14].

First, a systematic mapping review synthesizes existing literature on smart home-facilitated harm, resulting in a preliminary map. This method was selected because it aims to synthesize a research area and visually represent its results in a systematic map [13], aligning with the study’s aim. This map informs the second part: a workshop with

¹ <https://www.statista.com/forecasts/887636/penetration-rate-of-smart-homes-in-the-world>, last visited May, 2025.

smart home users who discuss and enhance the map. The second part can be seen as combining brainstorming and focus group techniques [14]. This is appropriate when research aims to understand ideas and general opinions about some products/devices [14], which also aligns with the study's aim. The subsections will detail both research parts and note some limitations.

2.1 Systematic Mapping Study

A systematic mapping study aims to provide a broad overview of a research area [13]. The outcome of such a review is a systematic map that contributes to a better understanding of the research area and supports future research efforts [13]. The present study follows the process of conducting a systematic mapping review as suggested [13]. Based on the RQ defined above, a search string was formulated that combines the study's two key concepts, harm and smart homes. 246 published, peer-reviewed articles in English were extracted from an appropriate database (EBSCO Discovery Service) and put in a reference handling system. Upon skimming titles, keywords, and abstracts, 123 were excluded as they were not specific to smart homes (exclusion criterion). Skimming the full texts of the remaining articles, 48 results were included in the study and subject to thematic analysis [13], as they mentioned a particular form of harm (inclusion criterion).

2.2 Workshop with Smart Home Users

The workshop aimed to build on the preliminary map produced in the systematic review and provided a starting point for the workshop participants' discussions. The workshop had four phases. The first was an introductory phase where the study and aim of the workshop were shared. Using a visual aid, the systematic map was projected on a whiteboard to introduce and define the categories. Lastly, the participants self-reported the years they had lived in smart homes and their types of smart devices (information in Table 1). The second phase put the participants in random pairs. There were three pairs and one group of three. The instruction in this configuration was to discuss the preliminary map and produce a mind map with revisions or additions. They were given post-its, pens, and big blank paper to document their maps. They were asked to include the categories they had been introduced to as they saw fit. When the groups finished, the third phase had them join with another group, creating two larger groups of four and five participants, respectively. They were asked to share the mind maps they had developed in the second phase and to merge them into one. If the groups had identified similar harms, the post-its were placed on each other or rewritten on a new one to transfer all the data appropriately. This procedure was also conducted in the fourth phase with one large group to produce a single map with all the identified harms. The harms did not have to be first-hand accounts and could include fears and anxieties about smart home living. The fifth and final phase was a wrap-up during which the facilitator took observational notes to capture the participants' reflections. The participants were asked about the workshop they had just undergone and what they thought of the map they had co-produced. The author translated and digitized the final map, depicted in Fig. 3. The workshop was in Swedish and lasted about 2 h.

Recruitment and Participants. For the workshop session, smart home users needed to be recruited. Due to time restraints, the selected method was to recruit participants through street intercepts. Three different places in Malmö, Sweden, were visited, and pedestrians were stopped and asked whether they were interested in participating in a workshop. For three days at the same time (6 pm), one location was visited: the first location was adjacent to the central station, the second was in a commercial area (approx. 7 km away from the central station), the third was near a residential area (approx. 7 km away from the central station and approx. 5 km away from the commercial area). The criteria to be included in the study were to have been living with two or more smart home devices for at least one year. After asking these questions and receiving affirmative answers, they were invited to the workshop later in the same week. Phone numbers were exchanged for details and other correspondence. The criteria were set for two reasons. first, harms and negative impacts of smart homes might not be obvious and become more apparent over time. Second, the study is not device-specific and considers smart homes to be a heterogeneous system. Therefore, having two or more devices was preferred. however, recruitment was difficult, and some participants did not meet one of the two criteria; see Table 1 for information about the participants. Only people walking alone or with children were approached to ensure that no participants knew each other or the facilitator (principal author).

Table 1. Self-reported information about participants in the workshop.

#	Experience	IoT devices
1	1 year	Smart virtual assistant, smart tv
2	2 + years	Smart virtual assistant, smart tv, smart doorbell
3	5 + years	Smart garage door, smart thermostat, smart tv, smart lawn mower, smart security camera, smart irrigation system
4	2–3 years	Smart locks, smart security camera, Smart TV
5	3 years	Smart fire alarm, smart security camera, Smart TV
6	2 years	Smart TV, smart security camera, Smart virtual assistant
7	3 + years	Smart TV, smart vacuum, smart lights
8	1 year	Smart virtual assistant
9	3 + years	Smart virtual assistant, smart TV, smart lights, smart vacuum, smart fire alarm

2.3 Limitations

Some limitations of the study are worth noting. Further analyses are needed to determine smart homes’ role in causing harm conclusively. The study aimed to overview the area and support future research efforts, aligned with the aim of systematic mapping studies that favor breadth over depth [13]. Thus, a wide range of articles was included in the

review. In Fig. 2, harms are shown as independent but should be viewed as layered with interconnected elements. Regarding the workshop, a small sample of smart home users was included, which only speaks to their experiences. A larger and diverse participant pool could have enhanced the map's comprehensiveness. For example, by including participants who live with children or aging parents, two common user groups of smart homes [5]. Similarly, the map's non-device-specific nature needs mentioning. Although use cases mentioning specific functionality were mentioned (see Fig. 3), including contextual information such as the participants' living situations and specific device types related to the harms could have provided more nuance to the results. Approaching the smart home as a unified system with undefined characteristics prompted wide-ranging results. Despite these limitations, the map depicted in Fig. 3, underpinned by previous literature and smart home users' experiences, supports the aim of the study: to provide an overview of the subject matter; smart home imaginaries and experiences of harm play an essential role in designing and developing smart home devices [15].

3 Results and Discussion

This section starts by reviewing and discussing the results from the systematic mapping study, presented in the preliminary map (Fig. 2). It then reviews and discusses the results from the workshop, and the final map is presented (Fig. 3).

3.1 Results from Systematic Mapping Study

Figure 2 depicts a preliminary map of eight harms facilitated by smart homes identified in the study. Some references mentioned more than one harm. The following is an initial review and discussion of some of the results underpinning the categories.

Aligned with previous reviews [4], domestic violence was a prominent category of harm facilitated by smart homes. The home has long been a place of violence [28], and the smart home exacerbates gender-based violence in alarming ways [33]. Smart homes have been seen to incite conflict due to the increased ability to surveil spouses, revealing conflict-provoking information [51]. Although there are benevolent motivations for surveillance, such as ensuring a spouse is safe, discussions submit to the correlation between spousal surveillance and in-person abuse [3]. Therefore, it is necessary to differentiate benign forms of spousal surveillance from harmful ones, in which the latter should be legislated as variations of domestic abuse [33].

Another category aligned with previous research is security breaches caused by external actors with malicious intent [47]. For example, when actors unfamiliar to the occupants manipulate devices to open a window, unlock a door [35], or cause harm remotely through malware attacks [47]. This category also includes harms related to hacking and unlawful retrieval of personal information. The experience of "being hacked" has reportedly generated great uncertainty for users [43]. Yet, users are commonly identified as the weakest link in security-related harm [47]. Although users have not been hacked, the anxiety related to, for example, interpreting unusual device behavior as signs of being hacked causes users stress [43]. The stress caused by the uncertainty of whether one has

been hacked deserves further attention. Also relevant to the smart home is the investments in smart city infrastructure, to which the smart home belongs. Security breaches can therefore be found on different layers, locally related to the smart home and city levels. Intruders could access smart homes from the smart city infrastructure [41]. The interplay between the smart home and smart city systems is apparent, and security vulnerabilities are acknowledged not as isolated risks experienced solely by smart homes [48]; security breaches could occur in both directions, from the smart home to the smart city as well as from the city to the home.

The results identified intrusive surveillance of children as a category of harm. Due to skewed power dynamics between children and their caregivers, surveillance can become controlling and impede children’s rights to autonomy while under the guise of good parenting [18]. The balance between care and control is both a personal decision for caregivers and a societal trajectory of what we consider to be acceptable parenting tactics [18]. Ensuring smart home devices are child-proofed and unhazardous was also a kind of harm related to this category [49].

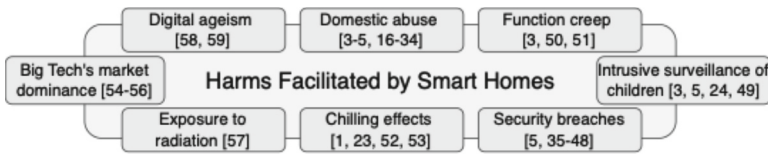


Fig. 2. A preliminary map of smart home-facilitated harm. References in brackets.

One result discussed the harms of increased radiation due to smart home infrastructure [57]. The reported impacts on smart home occupants were significant, and additional research into related harms, particularly to users’ health, is timely.

Substantial investments in smart nursing homes have been made, but their algorithmic harms to older users are poorly understood [58]. Smart homes designed for this user type risk replacing human care with data collecting and notification systems [59], and more attention needs to be paid to the harms of digital ageism: age-related bias in technology, such as exclusion from design and implementation processes [58].

Big tech companies’ hold on the smart home market [54] was identified as a harm. This was mainly discussed through the lens of surveillance capitalism [55] regarding how large conglomerates normalize surveillance and leverage their monopoly-like positions to stifle fair market competition [54]. As a result of their unruly access to personal data, societal and democratic harms are also discussed [54]. There are also differences between commercial products, which feed big tech’s grip on smart home markets, and the ability to develop and configure smart homes through DIY methods [56]. In DIY smart homes, users must be tech-savvy and understand adaptive learning and data mining techniques (common for IoT devices powered by artificial intelligence) [56]. The difference between harm facilitated by DIY smart homes compared to smart homes developed by big tech is unclear.

The last two categories relate mainly to domestic surveillance, a potential overarching category of smart home-facilitated harm. Domestic surveillance contributes to the platformization of the home [55], which, phenomenologically, can lead to harms related

to chilling effects; as users are tethered to the digitalized environment, they become modulated and disciplined by the smart home infrastructure [52]. Due to the level of surveillance and the chilling effects, users may adapt or suppress their behavior, which impedes their experience of being at home. Another category related to domestic surveillance is function creep, which occurs when devices are purchased, for example, to care for one's pet (pet tech) [50] and subsequently used for unintended purposes such as to facilitate unethical or intrusive forms of surveillance [3, 51]. An example found in the results is domestic violence being facilitated by pet tech [50], which is a change in intention from benign use to a harmful one. Function creep in smart homes is poorly understood and requires more thorough research.

3.2 Results from Workshop with Smart Home Users

All identified categories in the mapping study were included in the participant-developed map. Some were merged into larger categories or divided into smaller ones. For instance, digital ageism fell under the broader category of biased training data, which included gender and global north biases. Conversely, participants identified various types of security breaches. A common concern was the sensitivity of data collected from homes and the potential devastation of data extraction. From the observational notes, one participant explained: "First, I didn't think we would be able to think of so many, but once we started, it was like, woah, there are so many". Another participant echoed similar sentiments: "Once we got going, it was like there are so many potential ways this can be used against us". Given the number of categories and space constraints, all harms are not commented on in the present section. Below is a discussion regarding some categories the participants identified as harms facilitated by smart homes.

Unethical surveillance had the most subcategories. The mapping study revealed domestic surveillance as a potential overarching theme, including function creep and chilling effects. Participants viewed chilling effects as separate from unethical surveillance. Function creep and domestic violence were subcategories and also connected. Whether domestic violence is considered a form of function creep depends on the device's purchase intent. In severe cases, IoT devices may be acquired with abusive intentions. It indicates function creep if purchased for other uses but later used for control and violence. Such nuances require careful examination of IoT devices' role in domestic abuse and remains a distinct category of harm facilitated by smart homes.

Various relationships and power dynamics were identified under unethical surveillance. The mapping study revealed imbalanced dynamics between children and parents. Workshop participants noted further instances of skewed power between the surveilled and surveillers, like government surveillance of citizens. These relationships warrant deeper examination to ensure ethical surveillance practices.

A new category introduced by the participants, absent in the mapping study, was the environmental harms of smart homes. They noted a potential net negative impact linked to resource-heavy infrastructure, like server halls and training large models, poor waste management, and unsustainable consumption behavior. The inability to opt out of smart homes, increasingly seen as the norm, was another category added and was related to environmental harms. This was connected to an increased cost of living and how smart homes become resource-heavy, such as an increasing need for electricity. Safety and

security becoming a luxury good was also mentioned, exacerbating class differences and dividing society. In this sense, an overarching category of environmental harms of smart homes could include both ecological harms and harms to society at large; smart home-facilitated harm might impact people beyond the smart home users themselves.

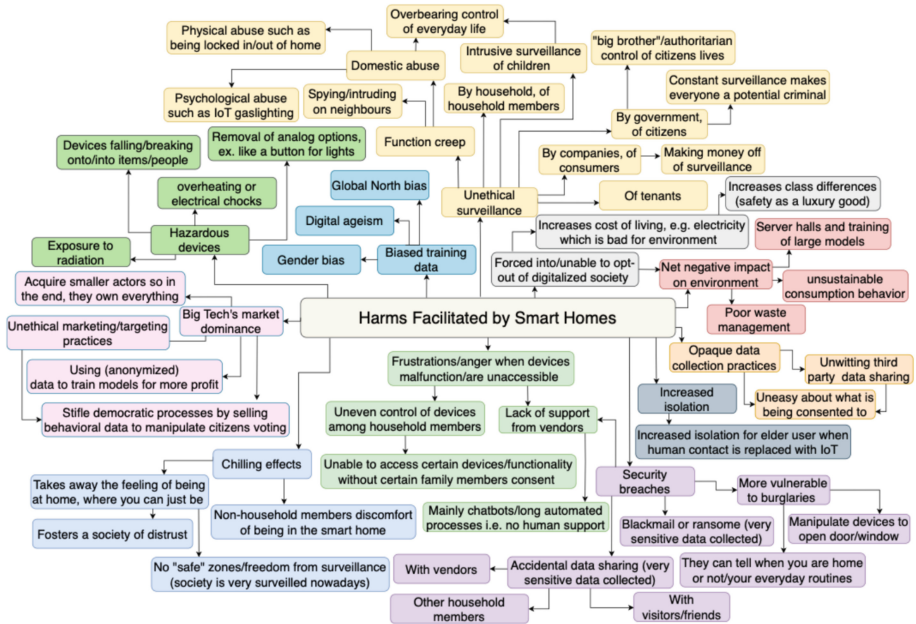


Fig. 3. Final map of smart home-facilitated harms. Colors are meant to simplify reading.

The mapping study identified big tech’s market dominance as a harm category, mentioning related societal and democratic harms. Workshop participants emphasized this connection, showing how big tech dominates markets through the acquisition of competitors, their ability to manipulate users via the large quantity and high variety of behavioral data collected from smart home devices, and the selling or re-purposing of such valuable data to third parties (including political campaigns and data brokers). The participants connected unethical marketing and targeting practices to democratic harms, as big tech can manipulate citizens’ voting behavior. Like the harms related to security breaches, the amount of sensitive and personal data that can be collected from smart homes is unique, compared to other public IoT environments. In a smart airport, for example, chilling effects occur naturally, and people don’t divulge intimate aspects of their personhood. With access to smart home data, more precise targeting can be achieved to sway users and influence their behavior. Further consideration should be paid to the home as a necessary free zone from behavioral suppression and the relationship between big tech, smart homes, and citizens’ right to freedom from surveillance.

Notably, although privacy is considered a common concern related to smart homes [11], it was not mentioned in either part of the research process. One explanation could be its embedded inclusion in other categories. For example, unethical surveillance can be

understood as a kind of privacy-violating harm. More research is required to understand the correlation between the two concepts, surveillance and privacy, and whether privacy breaches are caused by unethical surveillance or whether it is a standalone concern, as previous research suggests.

4 Conclusion and Future Work

This study aimed to overview smart home-facilitated harms. A systematic mapping study and workshop with nine smart home users identified several types of harms, contributing to understanding the wide range of harms smart homes may facilitate. Although more research is needed, the final map (Fig. 3) provides an overview of the kinds of harms smart homes can facilitate, addressing the study's stated RQ. With the rising popularity of smart homes, continuing to examine the harms they might facilitate is vital. Both in terms of educating users about the risks associated with smart homes and developing policy to mitigate the societal and democratic harms identified in this study, for example, environmental degradation and market monopolization. Such harms are not isolated, and their implications reach far beyond the boundaries of the home. Three avenues for future work follow. First, reiterating the research process to refine the categories and their relationships through additional focus groups and literature reviews with selected keywords. Second, analyzing how smart homes specifically facilitate harms by defining use cases, IoT device capabilities, descriptive scenarios, and incentive structures. Engaging diverse stakeholders, including law and policymakers, is crucial, especially for harms linked to domestic abuse and security breaches, which necessitate legal input. Similarly, deeper analyses regarding the specific impact of smart home devices' environmental harm are necessary, following the devices' lifespan from inception to disposal. Lastly, recognizing smart homes as part of a larger data ecosystem and examining the interplay between smart homes and smart cities. Exploring the ethical similarities and differences between these distinct, yet related environments is worthwhile and could reveal the trajectory that IoT investments put us on as a society. For example, related to the harms of big tech's market dominance. Although smart homes generate highly sensitive and personal data compared to public IoT environments, they should not be viewed as a separate IoT context. This is especially relevant regarding security breaches and data protection. The impact of deploying data-driven solutions within the home could have consequences potentially affecting the data security the smart city, as discussed in the results of the study. If governments and municipalities envision a fully integrated IoT life, then omitting the smart home and failing to acknowledge it as an integral part of the development is rather irresponsible. Assuming democratic and societal responsibility demands a further exploration of smart home ethics and their far-reaching harms, impacting not only its users but society at large.

References

1. Lupton, D.: The Internet of Things: Social dimensions. *Sociology Compass* **14**(4) (2020)
2. Maalsen, S.: Revising the smart home as assemblage. *Hous. Stud.* **35**(9), 1–16 (2019)

3. Dereymaeker, J., et al.: Smarter homes, smarter surveillance? Exploring intimate surveillance practices in modern day households. *New Media & Society* (2024)
4. Afrouz, R.: The nature, patterns and consequences of technology-facilitated domestic abuse: a scoping review. *Trauma Violence Abuse* **24**(2), 913–927 (2023)
5. Olabode, S., et al.: Complex online harms and the smart home: A scoping review. *Future Generation Computer Systems* **149** (2023)
6. Stucke, M.E., Ezrachi, A.: How digital assistants can harm our economy, privacy, and democracy. *Berkeley Technol. Law J.* **32**(3), 1239–1300 (2017)
7. Hammi, B., et al.: Survey on smart homes: Vulnerabilities, risks, and countermeasures. *Comput. Secur.* **117**(2), 102677 (2022)
8. Seymour, W., et al.: Strangers in the room: unpacking perceptions of ‘smartness’ and related ethical concerns in the home. In: *ACM Designing Interactive Systems Conference* (2020)
9. He, W., et al.: When smart devices are stupid: Negative experiences using home smart devices. In: *Security and Privacy Workshops*, pp. 150–155 (2019)
10. Ehrenberg, N., Keinonen, T.: The technology is enemy for me at the moment: How smart home technologies assert control beyond intent. In: *CHI Conference on Human Factors in Computing Systems*, pp. 1–11 (2021)
11. Zheng, S., et al.: User perceptions of smart home IoT privacy. In: *ACM on Human-Computer Interaction*. **2**, 1–20 (2018)
12. Wood, M.A., et al.: Inviting, affording and translating harm: Understanding the role of technological mediation in technology-facilitated violence. *Br. J. Criminol.* **63**(6), 1384–1404 (2022)
13. Petersen, K., et al.: Systematic mapping studies in software engineering. In: *12th International conference on evaluation and assessment in software engineering* (2008)
14. Shull, F., Singer, J., Sjøberg, D.I.: *Guide to Advanced Empirical Software Engineering*. Springer, London (2008)
15. Pereira, V.J., Hargreaves, T.: Are you thinking what I’m thinking? The role of professionals’ imaginaries in the development of smart home technologies. *Futures* (2024)
16. Leitão, R.: Anticipating smart home security and privacy threats with survivors of intimate partner abuse. In: *Designing Interactive Systems Conference*, pp. 527–539 (2019)
17. Winkle, K., Mulvihill, N.: Anticipating the use of robots in domestic abuse: a typology of robot facilitated abuse to support risk assessment and mitigation in human-robot interaction. In: *International Conference on Human-Robot Interaction*, pp. 781–790 (2024)
18. Sovacool, B., Furszyfer-Del Rio, D.D., Martiskainen, M.: Can Prosuming Become Perilous? Exploring Systems of Control and Domestic Abuse in the Smart Homes of the Future. *Frontiers in Energy Research* (2021)
19. Freed, D., et al.: Digital technologies and intimate partner violence: a qualitative analysis with multiple stakeholders. In: *ACM on Human-Computer Interaction*, pp. 1–22 (2017)
20. Lo, M.: A Domestic Violence Dystopia: Abuse via the Internet of Things and Remedies Under Current Law. *California Law Review* (2021)
21. Boethius, S., et al.: The double-edged sword - abused womens experiences of digital technology. *Eur. J. Soc. Work.* **26**(3), 506–518 (2023)
22. Shulruff, T.: Gender-based violence enabled by location technologies. *Int. J. Geograph. Info.* (2022)
23. Veel, K., et al.: HOMECTRL: an example of practice-based artistic research as a methodology in surveillance studies. *Surveill. Soc.* **20**(4), 379–389 (2022)
24. Sjølen, K.L.G., Veel, K.: The leaky home: conceptualising the remote control of smart technologies in the domestic sphere. *Theory Cult. Soc.* **42**(2), 65–80 (2025)
25. Backes, B.L., et al.: Leveraging video evidence in domestic violence misdemeanour cases: examination of case dispositions and case length. *Policing* **18**(9), 1–11 (2024)

26. Cavanagh, K., et al.: 'Remedial Work': Men's Strategic Responses to Their Violence Against Intimate Female Partners. *Sociology* (2001)
27. Doughty, K., Clark, J.: The role of telecare in the protection of individuals from domestic violenc. *J. Assist. Technol.* **3**(4), 44–48 (2009)
28. Slupska, J.: Safe at home: towards a feminist critique of cybersecurity. *St Antony's International Review* **15**(1), 83–100 (2019)
29. Brown, A., et al.: Safeguarding the "Internet of Things" for Victim-Survivors of Domestic and Family Violence: Anticipating Exploitative Use and Encouraging Safety-by-Design. *Violence Against Women* (2024)
30. McKay, D., Miller, C.: Standing in the way of control: a call to action to prevent abuse through better design of smart technologies. In: *CHI Conference on Human Factors in Computing Systems*, pp. 1–14 (2021)
31. Kouzani, A.Z.: Technological innovations for tackling domestic violence. *IEEE Access.* **11**, 91293–91311 (2023)
32. Brookfield, K., et al.: Technology-facilitated domestic abuse: an under-recognised safeguarding issue? *Br. J. Soc. Work.* **54**(1), 419–436 (2024)
33. Douglas, H., et al.: Law, Technology-facilitated Domestic and Family Violence: Women's Experiences. *British Journal of Criminology* (2019)
34. Lena, W.: The three harms of gendered technology. *Australas. J. Inf. Syst.* **24**, 1–9 (2020)
35. Adeeb Mansoor, A., et al.: Ontology-based classification and detection of the smart home automation rules conflicts. *IEEE Access.* **12**, 85072–85088 (2024)
36. Ahn, J., et al.: Design and implementation of hardware-based remote attestation for a secure internet of things. *Wireless Pers. Commun.* **114**(1), 295–327 (2020)
37. Buil-Gil, D., et al.: The digital harms of smart home devices: a systematic literature review. *Computers in Human Behavior* (2023)
38. Davis, B.D., et al.: Vulnerability studies and security postures of IoT devices: a smart home case study. *IEEE Internet Things J.* **7**(10), 10102–10110 (2020)
39. Jhuang, Y.-Y., et al.: GDPR Personal Privacy Security Mechanism for Smart Home System. *Electronics* (2023)
40. Mohd Nizam, O., et al.: A low-cost home security notification system using IoT and telegram bot: a design and implementation. *J. Comp. Res. Innov.* **7**(2), 327–337 (2022)
41. Qureshi, A., et al.: An efficient framework for security of internet-of-things devices against malicious software updates. *Computer Networks & Communications* (2024)
42. Rini, I., Elfizon, E.: Face recognition door security system and contactless bell. *Jurnal Teknik Elektro Indonesia.* **5**(1), 223–232 (2024)
43. Rostami, A., et al.: Being Hacked: Understanding Victims' Experiences of IoT Hacking. 18th Symposium on Usable Privacy and Security, pp. 613–631 (2022)
44. Vojković, G., et al.: IoT and smart home data breach risks from the perspective of data protection and information security law. *Business Systems Research* **11**(3) (2020)
45. Wijesundara, W.M.A.B., et al.: Security-enhanced firmware management scheme for smart home IoT devices using distributed ledger technologies. *Int. J. Inf. Secur.* **23**(3), 1–11 (2024)
46. Zhou, Y., et al.: Smart home cyberattack detection framework for sponsor incentive attacks. *IEEE Trans. Smart Grid.* **10**(2), 1916–1927 (2019)
47. Alshamsi, O., et al.: Towards securing smart homes: a systematic literature review of malware detection techniques and recommended prevention approach. *Information* **15**(10), 631 (2024)
48. Surjeet, D., et al.: Improving smart home surveillance through YOLO model with transfer learning and quantization for enhanced accuracy and efficiency. *Peer J. Comp. Sci.* (2024)
49. Sun, K., et al.: Child safety in the smart home: parents' perceptions, needs, and mitigation strategies. In *ACM on Human-Comp. Interact.* **5**, 1–41 (2021)
50. Harper, S., et al.: Security and privacy of pet technologies: actual risks vs user perception. *Frontiers in The Internet of Things* (2023)

51. Del Rio, D.D.F.: Smart but unfriendly: Connected home products as enablers of conflict. *Technology in Society* 68 (2022)
52. Burdon, M., Cohen, T.: Modulation harms and the google home. *Surveill. Soc.* **19**(2), 154–167 (2021)
53. Urquhart, L., et al.: Responsible domestic robotics: exploring ethical implications of robots in the home. In: *Journal of Information, Communication and Ethics in Society* **17**(2) (2019)
54. Hutchinson, C.S.: Curbing Big Tech’s IoT dominance. *European Competition Journal* (2022)
55. Pridmore, J., et al.: Intelligent personal assistants and the intercultural negotiations of dataveillance in platformed households. *Surveill. Soc.* **17**(1/2), 125–131 (2019)
56. Sanchez, A., Tercero, R.: Smart home technologies: uses and abuses. In: *Ninth Mexican International Conference on Artificial Intelligence*, pp. 97–102 (2010)
57. Reshna, R., Kheira Anissa Tabet, A.: An Appraisal Among Wired, Hybrid and Wireless Smart Homes to Mitigate Electromagnetic Radiation. *Frontiers in Built Environment* (2022)
58. Berridge, C., Grigorovich, A.: Algorithmic harms and digital ageism in the use of surveillance technologies in nursing homes. *Front. Sociol.* **7**, 957246 (2022)
59. Lussier, M., et al.: Integrating an Ambient Assisted Living monitoring system into clinical decision-making in home care: an embedded case study. *Gerontechnology* **19**(1) (2020)



Digital Identity and Control: How AI Replicas Challenge Performance Rights

Ahmed Qayyum and Stacy A. Doore^(✉) 

Department of Computer Science, Colby College, Waterville, ME, USA
maqayy26@colby.edu

Abstract. This paper explores the ethical challenges posed by the use of digital replicas in the entertainment industry, particularly in light of recent labor movements and rapid AI integration into production pipelines. Focusing on the reproduction of performer likenesses through artificial intelligence, the paper analyzes how these technologies challenge concepts of autonomy, informed consent, and artistic authenticity. Using recent examples from film and television, along with theoretical analysis grounded in deontological and utilitarian ethics, we examine how synthetic performances disrupt traditional models of compensation and control. The digital resurrection of deceased actors, the creation of synthetic performers, and the indefinite reuse of captured performances all raise unresolved legal and moral dilemmas. Drawing from the 2023 SAG-AFTRA labor negotiations and recent policy debates, the paper proposes a series of safeguards including continuous consent mechanisms, updated publicity rights laws, and greater industry transparency. We argue that responsible AI integration in entertainment must center human dignity and creative agency over technological expediency.

Keywords: artificial intelligence · digital replicas · entertainment industry · AI ethics · synthetic performance

1 Introduction

The entertainment industry stands at an ethical crossroads as artificial intelligence (AI) technologies are integrated into creative production processes. Recent developments in AI-generated content have sparked intense debates about the boundaries of creative control, performer autonomy, and artistic consent. The 2023 entertainment labor strikes cost entertainment producers over \$6 billion and highlighted how AI media integration threatens traditional definitions of authorship and performance rights [1, 2]. Among the most contentious innovations are digital replicas – sophisticated AI-generated performances that replicate a performer’s likeness, voice, and mannerisms with unprecedented accuracy. Unlike earlier forms of computer-generated imagery (CGI), today’s digital replicas pose unique challenges by enabling the indefinite reuse and manipulation of a performer’s digital identity, often without clear boundaries or ongoing consent.

Actors and creators argued that AI-generated synthetic performances threaten fundamental rights and economic fairness by potentially replacing human talent with endlessly reproducible digital likenesses.

This paper examines the ethical implications of AI-generated production in media and entertainment, focusing particularly on performer autonomy, intellectual property rights, and the consent of stakeholders whose creative work and likenesses are being digitally replicated. Cases such as the digital resurrection of deceased actors and depictions of ordinary people in speculative dystopian series being exploited for entertainment and profit underscore the urgent need for industry-wide ethical frameworks and legal safeguards. Through analysis of current industry practices, legal frameworks, and ethical guidelines, we argue these technologies challenge long-held norms about the value of creative labor [3]. We further contend that existing legal and regulatory structures are insufficiently equipped to address the novel challenges posed by AI use in filmmaking.

Drawing on both fictional narratives and real-world implementations, we investigate how the industry grapples with these challenges and propose potential solutions for protecting creative rights while fostering technological innovation, to safeguard performer rights and maintain the integrity of creative expression. Through analysis of current industry practices, legal frameworks, and ethical guidelines, we argue how these technologies challenge long-held norms about the value of creative labor and the concept of intellectual property and likeness brand ownership in the entertainment sector [4].

The stakes in this discussion extend beyond individual rights to the very future of creative expression in media; as AI systems become more sophisticated, the entertainment industry must find ways to balance technological advancement with the protection of human creativity and artistic integrity. This paper contributes to this dialogue by examining the ethical frameworks necessary for responsible AI integration in entertainment production.

2 Background

Following the entertainment industry's rapid embrace of artificial intelligence (AI), understanding the technical evolution and current capabilities of these technologies becomes crucial for analyzing their ethical implications. This technological transformation has introduced new tools and methods that fundamentally alter traditional creative processes and require careful examination of their impact on the industry's future. While early applications of AI focused primarily on computer-generated imagery (CGI), contemporary AI systems span the entire production pipeline, from visual effects and character animation to analyzing performer metrics for casting decisions and handling post-production tasks like color grading and sound mixing [5]. The advent of deep learning and neural networks marked another pivotal shift, enabling AI to become an active participant in the creative process, particularly evident in performance capture. Traditional CGI required extensive human intervention, but modern AI systems can automatically generate complex visual elements and entire synthetic performances. Kim [6] documents how Hollywood studios increasingly employ AI

for tasks ranging from background generation to complete character animation, while the South Korean film industry has integrated AI into visualization and shot planning. Major studios have reported efficiency gains of up to 40% in certain production phases by means of AI implementation [6]. Such technological developments have profound implications for the entertainment industry's structure and operations. AI tools are reshaping traditional job roles, particularly in areas like visual effects, script development, and background character generation.

2.1 Ethical Issues

While these advances offer potential cost savings and creative possibilities, they also raise significant concerns about job displacement and fair compensation for individuals who are veterans in the media industry [7]. Entry-level positions that serve as crucial stepping stones for creative talent face replacement through digital automation, disrupting established career pathways. AI's rapid transformation of content creation and distribution models have also reshaped traditional production workflows, significantly compressing schedules and altering employment patterns without clear adjustments to compensation or job security [8]. The industry faces unprecedented challenges in balancing technological efficiency with the preservation of creative professions and artistic integrity in an AI-driven industry [4]. The concentration of technological power in the hands of a small number of major studios also raises concerns about market diversity and creative independence [6]. This could lead to a homogenization of content as AI systems, trained on existing successful works, replicate proven formulas rather than encourage innovative storytelling [9]. Additionally, the global nature of AI implementation complicates standardization and regulation, creating disparities in creative protections and market accessibility across different regions.

2.2 Copy Without an Original

Perhaps the most ethically complex AI advancement, digital replica technology, enables the creation of synthetic performances that can replicate an actor's likeness, voice, and mannerisms [10]. Generating new performances from existing performance data creates a "simulacrum" - a copy without an original - challenging traditional notions of performance authenticity and artistic ownership [11]. The 2023 labor strikes brought these concerns to the forefront of industry discourse. A key point of contention centered on what the final agreement termed "synthetic performers" - AI-generated performances not based on specific actors - versus "digital replicas" of actual performers. This technological advancement directly leads to ethical debates surrounding authenticity, consent, and creative autonomy [3]. Unlike AI's impact on technical production roles, digital replicas specifically transform the bargaining position of actors, who must now consider how their digital likenesses might be used long after their initial performance capture, creating new power imbalances between studios and performers in contract

negotiations. These discussions are seen in both fictional narratives and industry practices illustrating the complex relationships between performer identity, creative control, and commercial exploitation in the AI age.

3 Fictional Narratives and Industry Practices

The entertainment industry's grappling with AI has now become a prominent theme in contemporary media, with fictional narratives often preceding and predicting real-world developments. These cultural works serve not only as entertainment but as crucial commentary on the ethical implications of AI in creative industries to shine a spotlight on the ethical issues discussed above.

Recent television and film productions have demonstrated an uncanny ability to anticipate the ethical challenges of AI in entertainment. Black Mirror's "Joan is Awful" episode presents a particularly prescient examination of likeness rights and consent in the age of AI [12]. The episode, which depicts a woman discovering her life being adapted into an AI-generated streaming series without her meaningful consent. The story centers on the exploitation of personal data by a tech giant to produce inexpensive, hyper-realistic content. Similarly, "Rachel, Jack and Ashley Too" from the same series explores the commodification of performer identity through AI replication, following the story of a pop star whose consciousness is uploaded into a robotic doll, which raises concerns about creative authenticity and performer exploitation [13].

"The Congress" stands as a striking example of media foresight regarding AI's impact on entertainment. And an exploration of digital replica ethics [14]. The film follows Robin Wright, playing a version of herself, who sells her digital likeness to a studio, allowing them to create new performances without her involvement. What seemed like science fiction in 2013 now reads as a prescient warning about current industry practices. The film explores themes of performer autonomy, consent, and the commodification of artistic identity that directly parallel current industry debates about AI-generated content and digital replicas [15].

Beyond depictions of these ethical issues in media, the industry's actual engagement with AI technology has also generated significant controversy, particularly regarding posthumous reproduction of performer identities. The "Roadrunner" documentary's use of AI to generate Anthony Bourdain's voice without explicit prior consent sparked intense debate about the ethics of digital recreation [16], where the gap between technological capability and ethical considerations in AI implementations was highlighted to audiences. The digital recreation of the deceased Peter Dinklage in "Rogue One" [17] raised questions about posthumous performance rights and the boundaries of estate consent. While Dinklage's estate approved the performance recreation, the case sparked conversations about whether such approval adequately addresses the ethical implications of digitally resurrecting performers [18].

Recent films have used AI technologies to create younger versions of actors such as Harrison Ford, Tom Hanks, and, ironically Robin Wright, raising questions about control of one's historical likeness. The 2023 film "Indiana Jones and

the Dial of Destiny” employed extensive AI-powered de-aging of Harrison Ford for its 25-minute opening sequence. This technique drew criticism for falling into an uncanny valley - where the de-aged face appeared young but was animated with the movements and physical presence of the older actor, creating a disconnect that many viewers found distracting [19]. These AI methods operate in a gray area between enhancement and replication, challenging traditional understandings of performance authenticity. The industry has yet to establish clear ethical guidelines for when and how such technology should be employed, particularly regarding informed consent for future uses that may not be technologically possible at the time of initial agreement.

More controversially, the 2025 Polish biopic “Putin” marked the first instance where AI was used to digitally transform lead actor Sławomir Sobala’s appearance to resemble the Russian president throughout an entire feature film, raising profound questions about likeness rights when portraying living public figures without their consent, especially for politically sensitive depictions [20]. The industry has yet to establish clear ethical guidelines for how such technology should be employed, particularly regarding informed consent for future uses that may not be technologically possible at the time of the initial agreement.

4 Ethical Analysis

Traditional acting theory emphasizes the importance of lived experience - the performer’s conscious presence in space and time, their physical embodiment of emotion, and their dynamic interaction with other performers and the environment. When a digital replica is created, there is a separation of the external manifestations of performance from the actor’s embodied consciousness, raising questions about whether an authentic performance requires a conscious experience or if a perfect simulation of only the external elements is sufficient [3]. In traditional performances, audiences understand that a human actor is consciously crafting a role. With digital replicas replacing actors, the layers of deception multiply with little clarity for the audience as to the human agency in the performance. [7]

4.1 Performer Autonomy and Continuous Consent

From a Kantian deontological perspective, which emphasizes individual autonomy and human dignity, this fundamental separation violates core ethical principles. The categorical imperative – to treat humans as ends in themselves rather than merely means – reveals critical issues in digital replica deployment [18]. When AI systems generate content based on existing performances, they violate the moral rights of creators to their intellectual expression. Infinitely reproducing and manipulating an actor’s likeness without ongoing consent reduces their autonomy to mere instrumental value. This becomes particularly problematic when considering Kant’s emphasis on authenticity and truthfulness - AI-generated performances, by their nature, involve a form of deception about

their origins and creation process. This technological shift risks turning actors into mere “source material” for digital duplicates rather than continuing creative contributors, fundamentally altering their professional identity, agency, and economic value in the industry.

This ethical complexity deepens when considering the temporal nature of consent involved within the creation of digital replicas. Unlike traditional film performances, where an actor’s work remains fixed in its original context, AI technology enables the endless manipulation and repurposing of their likeness. This raises fundamental concerns about the nature of performance consent: can an actor truly provide informed consent for future uses of their digital replica that may not even be technologically possible at the time of the original agreement? As Curren argues [3], this creates an unprecedented ethical dilemma where performers must make decisions about their professional future without fully understanding the potential implications. Once an actor’s likeness is captured and digitized, controlling its future use becomes increasingly complex, and traditional contracts, designed for a pre-AI era, fail to address scenarios where an actor’s digital double could be repurposed indefinitely. The right of publicity laws, while offering some protection, haven’t evolved quickly enough to address these new forms of digital reproduction. The creation and use of digital replicas without continuous consent, then, fundamentally violates the principle of respect for persons’ autonomy. Preminger and Kugler [18] conclude that this practice effectively commodifies human identity, reducing performers to digital assets that can be manipulated at will. The inability of performers to maintain control over their digital likeness throughout its use represents a fundamental violation of their right to self-determination under deontological ethics.

4.2 Economic and Professional Implications

Moreover, the power dynamics between performers and studios add another layer of ethical concern in the digital replica landscape. Wang notes [7] actors often face economic pressure to agree to broad terms regarding their digital likeness, effectively forcing them to choose between current employment and future autonomy. There is an added coercive element which undermines the fundamental principle of informed consent, as actors might feel compelled to sign away rights to their digital identity simply to remain competitive in the industry [8]. Performers also face the unprecedented challenge of competing against their own digital replicas, undercutting their own market value and career longevity.

The imbalance is particularly pronounced for background performers and non-featured actors who typically possess minimal negotiating leverage. The current model enables studios to profit from captured performances without fairly compensating the original performers, establishing a troubling precedent where an actor’s digital twin can continue generating revenue long after the initial compensation agreement [4]. During the production of Disney and Marvel Studio’s WandaVision [21], background actors were reportedly directed to undergo full-body scans without clear explanation of how their digital likenesses would be utilized [22]. According to statements from actors, these scans involved

capturing various facial expressions and body positions, yet the actors were not informed about the potential future use of their digital replicas, nor were they compensated beyond their standard day's pay.

From a utilitarian perspective, these immediate economic advantages for studios create complex ethical tradeoffs. While the technology offers potential benefits like increased content production efficiency, reduced costs, and greater accessibility to entertainment content, the long-term consequences present serious concerns. The displacement of creative workers creates ripple effects throughout society - not just economic hardship for individuals, but a broader decline of creative professions. Bender notes [9] that this could eventually lead to reduced diversity in storytelling, as AI systems trained on existing works tend to replicate rather than innovate, and so the audience's viewing experience might actually deteriorate in the long run.

The transformation of actor identity into a licensable digital asset introduces further ethical implications regarding compensation models. Traditional residual payments, designed to ensure performers continue benefiting from successful work, become obsolete when a studio can create new performances from an existing digital replica without the actor's involvement. The economic value of an actor's career-defining characteristics—their expressions, mannerisms, and physical presence—becomes detached from the performer themselves, creating a “residuals crisis.” [4]. This creates a ‘consent paradox’ where individuals must make decisions about their digital rights and artistic legacy without truly understanding potential future implications [10].

The international dimensions of digital replica usage present additional challenges. Different markets and cultural contexts approach digital likeness rights based on their own industry interests, creating potential inequities in how performers are protected across various regions [6]. Global professional standards become crucial as productions might exploit regulatory gaps to create digital replicas in jurisdictions with fewer protections, effectively circumventing stronger regulations elsewhere.

4.3 Audience Reception

The ontological status of digital replicas raises further philosophical questions. When we create a digital performance by combining elements of an actor's appearance, voice, and mannerisms with new movements and contexts at what point does it cease to be authentically “their” performance? Each iteration or modification of a digital replica represents a replacement of some aspect of the original performance, which challenges our understanding of authenticity and ownership in creative expression. We can replicate the physical aspects of performance perfectly, but something ineffable about human consciousness remains beyond our ability to digitize.

This leads us to question whether truly authentic performance requires consciousness, or if perfect simulation of the external elements is sufficient. This becomes particularly intricate when taking into consideration performances that merge multiple temporal versions of an actor through de-aging technology, as

with Tom Hanks in the movie “Here” [23], or posthumous performances that combine elements from old performances to create something entirely new.

Thus, a comprehensive utilitarian analysis must consider how digital replicas fundamentally alter audience experience through forms of deception. The knowledge that performances might be artificially generated could erode audience trust and emotional connection to creative works, such as in the case of the Roadrunner documentary that utilized Anthony Bourdain’s voice to simulate him reading private emails and passages he never spoke aloud, without explicit disclosure to the audience [24]. More recently, as seen with the Oscar winning movie “The Brutalist” [25], even when AI is used for seemingly minor adjustments like accent refinement rather than complete performance creation, questions arise within audiences about transparency, authenticity, and where to draw the line between technical enhancement and fundamental alteration of an actor’s work. Cases like these could intrinsically alter the relationship between audiences and entertainment, and end up reducing the overall social utility audiences derive from filmgoing [7].

Professional organizations have begun specifically addressing the issues digital replicas pose, with SAG-AFTRA securing groundbreaking protections in their 2023 agreement. These protections require informed consent for digital replica creation and usage, fair compensation for digital replica performances, and transparency in how likenesses will be used. Establishing comprehensive protections requires completely rethinking how we value and protect an actor’s likeness and performance rights in an era where those assets can be infinitely reproduced and modified [2].

5 Recommendations

While AI technology offers clear entertainment industry benefits, we argue its current deployment prioritizes efficiency, cost savings, and deception over human dignity, creative rights, and long-term societal welfare. We suggest there are several key policy recommendations for protecting creative labor rights while still enabling creative AI advancement.

5.1 Contracts

First and foremost, contracts governing the use of performer likenesses must be substantially restructured to address the ‘perpetual’ nature of digital replicas. Preminger and Kugler [18] declare that traditional one-time consent agreements are insufficient in an era where AI can indefinitely reproduce and modify performances. New contractual frameworks should include explicit provisions for continuous consent, requiring studios to obtain renewed permission for each novel use of a digital replica. These agreements should thus detail specific approved uses, compensation structures, and clear limitations on how AI systems can manipulate or repurpose performer likenesses. Additionally, contracts should include ‘sunset clauses’ that require periodic review and renewal of consent

terms, ensuring that performers can maintain agency over their digital representations as technology evolves. The right to withdraw consent represents another crucial policy consideration. As AI capabilities advance, performers may face unforeseen or unexpected consequences from earlier agreements to use their digital likeness. Performers should retain the ability to revoke or modify consent for future uses of their digital replicas, particularly when new technological capabilities emerge that weren't contemplated or discussed in original agreements [2]. This policy would need to balance performer autonomy with studios' legitimate business interests, perhaps incorporating structured notice periods or proper compensation mechanisms for consent withdrawal. The right to withdraw could be structured similarly to existing image rights protections [2], but with specific provisions addressing the unique challenges of AI-generated content.

The right of publicity law, in particular, needs considerable revision to accommodate AI technologies. Current publicity rights laws fail to adequately protect performers from unauthorized AI reproductions of their likeness [3]. New legal frameworks should expand the scope of publicity rights to explicitly cover AI-generated content, including not just visual reproductions but also synthesized voices, mannerisms, and performance styles. This may involve creating a new classification of "digital performance rights" that specifically addresses AI replication and manipulation. Such legislation would need to carefully balance First Amendment considerations with performer protections, particularly in cases where AI-generated content serves artistic purposes.

5.2 Compensation

There is a need for new categories of residual payments for AI-generated content that uses actor likenesses or writer contributions, to ensure that creative professionals continue to benefit from their work even as it is digitally repurposed. Compensation policies must also evolve to ensure fair payment in an AI-augmented industry. Gordon-Levitt, for example, proposes a comprehensive restructuring of payment models, including establishing minimum compensation standards that apply regardless of whether AI tools are used in production [4]. These policies should also address compensation for the use of creative works in AI training data, perhaps establishing industry-wide standards for licensing creative content for AI development. Such policies would help prevent the devaluation of creative work while providing clear frameworks for fair compensation in a technological landscape that is constantly evolving.

This would help protect creative work while providing clear frameworks for fair compensation in an evolving technological landscape. Current publicity rights laws fail to adequately protect individuals from unauthorized AI reproductions of their likeness. New legal frameworks should expand publicity rights to cover AI-generated content, including visual reproductions, synthesized voices, mannerisms, and performance styles [3]. These "digital performance rights" would regulate AI replication, define fair use for AI training, and establish standards for attribution when generating new content based on existing works [26].

Aligning international law is another important consideration. As AI technology transcends national boundaries, differing legal frameworks may continue to create challenges for protecting creative rights globally. There is a need for international treaties or agreements that establish consistent standards for AI use in entertainment, similar to existing international copyright conventions [6]. Such frameworks could establish minimum standards for consent, compensation, and creative rights protection across jurisdictions and nations and may help limit production companies from actively seeking locations with weaker protections for creatives.

5.3 Transparency

The successful integration of AI in entertainment requires more than just policies and laws; it demands fundamental changes in how studios approach transparency and ethical oversight in their use of AI technologies. A comprehensive framework for industry accountability will help ensure that technological advancement doesn't come at the expense of artistic integrity. First, studios must implement transparent disclosure protocols regarding their use of AI in productions. This goes beyond simple acknowledgment of AI use to include detailed documentation of how AI systems are employed throughout the creative process. Drawing from the SAG-AFTRA agreements, studios should be required to maintain clear records of AI implementation, including specific uses of digital replicas, AI-generated content, and any manipulation of performer likenesses [2]. Through the SAG-AFTRA agreed recommendations, actors now have the right to bargain over future use of synthetic performers, and they also are to be fairly compensated for the time spent to create their own digital replicas. Moreover, We argue that this transparency should extend to audiences as well, with clear notification when AI has been used to generate or modify performances, ensuring that viewers can make informed decisions about the content they consume.

The establishment of dedicated ethics committees within entertainment organizations also represents another crucial step toward accountability. These committees, which should be composed of diverse stakeholders including creative professionals, technologists, and ethics experts, would be responsible for evaluating AI implementations against established ethical guidelines. To address this issue, Bender (2024) [9] suggests these committees could serve multiple functions: reviewing proposed AI uses for potential ethical concerns, establishing best practices for AI integration, and providing oversight of ongoing AI projects.

Furthermore, industry-wide standards for AI auditing and compliance need to be developed. This may include regular assessments of AI systems' impact on creative professionals, monitoring for potential biases or unfair practices in AI implementation, and ensuring compliance with any established ethical guidelines. Regular review processes should be implemented to evaluate how AI technologies affect different stakeholder groups within the industry [8]. These standards should also include mechanisms for addressing any grievances that arise and providing remedies or solutions when AI implementations violate established ethical principles or professional rights.

6 Conclusion

The entertainment industry must strike a balance between embracing AI innovation and protecting creative rights for the benefit of all. Implementing accountability measures like these would help ensure that the entertainment industry's adoption of AI technology proceeds in a fashion that respects both creative rights and artistic vision, all the while still attempting to foster innovation. The contributions of this paper provide a foundation for ethical creative sector AI integration. However, the successful implementation of these protections requires an active commitment from all stakeholders: studios, creative professionals, technology developers, and industry regulatory bodies.

References

1. Scherer, M.: The SAG-AFTRA strike is over, but the AI fight in Hollywood is just beginning. 2023, center for Democracy & Technology. <https://cdt.org/insights/the-sag-aftra-strike-is-over-but-the-ai-fight-in-hollywood-is-just-beginning/>
2. Shetler, K.: AI and consent: what the SAG-AFTRA and WGA agreements tell us about the future of generative AI (2024). https://scholarship.shu.edu/student_scholarship/1483/
3. Curren, A.: Digital replicas: harm caused by actors' digital twins and hope provided by the right of publicity. *Texas Law Rev.* **102**, 155 (2023). <https://texaslawreview.org/wp-content/uploads/2024/01/Curren.Printer.pdf>
4. Gordon-Levitt, J.: If artificial intelligence uses your work, it should pay you (2023). *Washington Post*. <https://www.washingtonpost.com/opinions/2023/07/26/joseph-gordon-levitt-artificial-intelligence-residuals/>
5. Luchen, F., Zhongwei, L.: ChatGPT begins: a reflection on the involvement of AI in the creation of film and television scripts. *Front. Art Res.* **5**(17) (2023). <https://doi.org/10.25236/FAR.2023.051701>
6. Kim, J.G.: Current use and issues of generative AI in the film industry. *J. Inf. Technol. Appl. Manag.* **31**(3), 181–192 (2024). <https://doi.org/10.21219/jitam.2024.31.3.181> <https://doi.org/10.21219/jitam.2024.31.3.181>
7. Wang, Y.: Synthetic realities in the digital age: navigating the opportunities and challenges of AI-generated content (2023), *arXiv Preprints*. <https://www.techrxiv.org/doi/full/10.36227/techrxiv.23968311.v1>
8. AbuMusab, S.: Generative AI and human labor: who is replaceable? *AI & SOCIETY*, pp. 1–3 (2023). <https://doi.org/10.1007/s00146-023-01773-3>
9. Bender, S.: Generative-AI, the media industries, and the disappearance of human creative labour. *Media Pract. Educ.* 1–18 (2024). <https://doi.org/10.1080/25741136.2024.2355597>
10. Yamazaki, A.: Digital replicas and democracy: issues raised by the Hollywood actors' strike. *Humanit. Soc. Sci. Commun.* **11**, 1661 (2024). <https://doi.org/10.1057/s41599-024-04204-w>
11. Baudrillard, J.: *Simulacra and Simulation*. University of Michigan Press (1994)
12. Brooker, C., Pankiw, A.: *Joan is Awful* (Black Mirror, season 6, episode 1), netflix (2023)
13. Brooker, C., Sewitsky, A.: *Ashley Too* (Black Mirror, season 5, episode 3), netflix (2019)

14. Folman, A.: The congress, pandora Filmproduktion (2013)
15. James, C.: The congress: the prescient film that foretold Hollywood's AI crisis, bBC (2023). <https://www.bbc.com/culture/article/20230717-how-2013-film-the-congress-predicted-hollywoods-current-ai-crisis>
16. Limbong, A.: AI brought anthony bourdain's voice back to life. should it have? nPR (2021). <https://www.npr.org/2021/07/16/1016838440/ai-brought-anthony-bourdains-voice-back-to-life-should-it-have>
17. Edwards, G.: Rogue one: A star wars story. lucasfilm, Walt Disney Motion Pictures (2016)
18. Preminger, A., Kugler, M.B.: The right of publicity can save actors from deep-fake armageddon. Berkeley Technol. Law J. **39**(2) (2024). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4563774
19. Davis, C.: Attention, Hollywood: De-aging isn't working, so please stop using it, variety (2023). <https://variety.com/2023/film/awards/indiana-jones-5-harrison-ford-de-aging-not-working-1235618698/>
20. Etemesi, P.: The upcoming putin biopic is a bad idea (and not for the reasons you think), movieWeb (2025). <https://movieweb.com/upcoming-putin-biopic-stars-artificial-intelligence/>
21. Schaeffer, J.: Wandavision, marvel Studios (2021)
22. NPR: Movie extras worry they'll be replaced by AI — Hollywood is already doing body scans (2023). <https://www.npr.org/2023/08/02/1190605685/movie-extras-worry-theyll-be-replaced-by-ai-hollywood-is-already-doing-body-scan>
23. Zemeckis, R., Roth, E.: "Here," Film, 2024, starring Tom Hanks, Robin Wright, Paul Bettany, Kelly Reilly. <https://www.imdb.com/title/tt18272208/>
24. Yang, M.: Anthony bourdain documentary sparks backlash for using AI to fake voice, the Guardian (2022). <https://www.theguardian.com/food/2021/jul/16/anthony-bourdain-documentary-ai-voiceover-roadrunner>
25. Corbet, B., Fastvold, M.: "The Brutalist," Film, 2024, starring Adrien Brody, Felicity Jones, Guy Pearce, Joe Alwyn, Raffey Cassidy, Stacy Martin, Emma Laird, Isaach de Bankolé, Alessandro Nivola. <https://www.imdb.com/title/tt8999762/>
26. Kavitha, L.: Copyright challenges in the artificial intelligence revolution: transforming the film industry from script to screen. Trinity Law Rev. 4(1), 1–8 (2023). https://dlwqtxts1xzle7.cloudfront.net/104027427/1_Copyright_Challenges_in_the_1_8-libre.pdf



Through the Educators' Lens: University Teachers' Perceptions of AI Integration in Higher Education

Ewa Duda^(✉) , Ewelina Młynarczyk-Karabin ,
and Julia Więckiewicz-Modrzewska

Maria Grzegorzewska University, Warsaw, Poland
{eduda, emlynarczyk, jwieckiewicz}@aps.edu.pl

Abstract. The integration of artificial intelligence (AI) and chatbot technologies in higher education has accelerated in recent years, prompting significant pedagogical and institutional shifts. Using a mixed-methods approach, including a survey containing both closed and open-ended questions, we explore university teachers' perceptions of chatbots as educational tools. The study focuses on teachers' usage patterns, attitudes, concerns and perceptions of the ethical issues related to the use of AI-based tools. The findings indicate that teachers are most familiar with ChatGPT and AI language tools, such as language translation and speech-to-text transcription, as well as online support tools to improve writing, mainly because they use these tools in their professional work. Teachers who specialised in pedagogy believed to a greater extent than academics in other social science disciplines that pupils, students and teachers should participate in training on using AI. The study also explored ethical issues related to teachers' use of chatbots. Almost half of the respondents expressed security and privacy concerns when using a chatbot; however, academics with a doctorate were less likely to have privacy concerns than those with a master's degree. Over seventy percent of teachers believe that chatbots can be used to manipulate information or mislead students. They also believe that chatbots could increase inequalities in education. Recommendations are offered for policy makers, developers, and institutions aiming to integrate chatbots in ways that uphold academic standards and enrich the teaching-learning experience.

Keywords: Artificial Intelligence · Chatbots · ChatGPT · Higher Education · University Teachers · Technology Enhanced Education

1 Introduction

Technological advances in artificial intelligence (AI) are increasingly influencing the reality we live in and covering more and more areas of it. Whether we like it or not, we should accept that it is becoming an integral part of our lives. Similarly, we should accept that AI is becoming an integral part of our lives, including higher education. This brings opportunities for AI development, as well as new and unforeseen challenges, including

ethical ones. Contemporary higher education institutions are facing the need to adapt to the growing role of AI tools, which are catalysing transformation across diverse sectors and profoundly impacting daily practices [1]. Within higher education, AI technologies, from adaptive learning platforms to automated grading systems, promise to enhance pedagogical efficiency, personalise student learning, and streamline administrative tasks. They also offer support teachers in their research work. At the same time, a number of ethical dilemmas have arisen concerning the impact of AI on academic integrity, educational quality, the nature of academic work, and the evolving role of academics within the educational process.

As institutions increasingly adopt AI-based tools, the perspective of those at the forefront of implementation, like academics, remains under-researched. Whilst extant research has focused predominantly on student outcomes, institutional benefits or ethical debates, this study shifts the focus to academics, whose insights are crucial to understanding the true impact of artificial intelligence on teaching and learning dynamics. University teachers are navigating a dual reality. On the one hand, artificial intelligence offers them opportunities to reduce administrative burdens, implement pedagogical innovations and meet diverse student needs, but it also raises concerns about student autonomy, pedagogical autonomy and the ethical implications of algorithmic bias. Tools such as plagiarism detectors and AI-based tutoring systems require teachers to reconsider traditional assessment methods and student engagement strategies. Moreover, the rapid evolution of AI requires constant professional adaptation, triggering debates about redefining the teachers' role in an increasingly automated environment [2].

This study explores how higher education staff perceive the integration of AI technologies into their teaching and research practices. Through an analysis of the data collected, it examines the opportunities academics identify, the challenges they face, and the broader implications for pedagogy. By centring the voices of university teachers, this research aims to increase knowledge about the possibilities for expanding institutional support frameworks and equitable AI adoption strategies that are aligned with pedagogical goals. Ultimately, the findings aim to bridge the gap between technological innovation and human-centred education, ensuring that AI serves as a tool for empowerment rather than disruption in academia.

The article begins with an overview of current applications of artificial intelligence in higher education and describes the current study in more detail, followed by an analysis of university teachers' experiences and concerns about their use of AI-based tools. It concludes with recommendations for fostering collaborative AI integration that prioritises pedagogical integrity and educator agency.

2 Applications of Artificial Intelligence in Higher Education

From the perspective of scientists, AI technology has the potential to become a valuable complement to the teaching process, especially in the context of individualising teaching, increasing work efficiency, and providing access to educational resources. However, its widespread use also raises a number of concerns, particularly regarding the inappropriate use of technology by students and its impact on traditional teaching methods and academic values, especially in the context of the wide possibilities offered by chatbots

[3]. The number and scope of research undertaken on the use of AI programs in education is still growing. Scientists emphasise that the potential of AI to improve cooperation and communication between the scientific community and educational stakeholders is enormous. The results of studies conducted by various researchers [4–6] indicate that AI can effectively analyse student data, formulate personalised recommendations, adjust the scope and pace of educational content, and consequently positively affect learning outcomes.

The use of technology in the so-called adaptive learning creates opportunities to improve the quality of education by adjusting the learning process to the individual needs of the student, preferences, interests, and progress achieved, thus providing feedback. AI is also used by academic staff, because it can free them from time-consuming administrative tasks. According to research conducted by Tareq Rasul and colleagues [7], the use of chatbots can help in assessing student work, formulating tests and analysing student results, allowing teachers to focus on more demanding aspects of teaching, such as mentoring and individual work with students. Furthermore, the use of AI gives academics the opportunity to use increasingly innovative assessment methods that cannot be effectively implemented using only a theoretical approach. In addition, chatbots can be used to quickly analyse large data sets, which allows for the improvement of administrative processes in education [8]. The adoption of AI in teaching environments might reshape the traditional role of an academic teacher in the learning process. As [7] argue, the growing capabilities of AI may reduce the importance of some teacher functions, such as imparting factual knowledge, and increase the importance of skills related to mentoring, critical thinking, and evaluating sources. However, this shift also risks weakening the authority of scientists and reducing their influence on student development, which requires new teaching strategies and ethical guidelines for using AI.

Empirical studies of scientists' attitudes toward AI reveal a mix of positive and negative reactions. Many scientists see the potential for chatbots to improve teaching and research, including tasks related to generating ideas, assisting with literature reviews, and improving the flow of scientific papers. On the other hand, concerns about academic integrity, the quality of AI-generated content, and disruption of traditional teaching models have led many researchers to argue for the need for clear ethical principles governing the use of chatbots in teaching and research [9]. Other researchers have also pointed to difficulties arising from the use of AI. [10, 11] indicate that chatbots provide information solely based on the database they have. This may lead to the formulation of not always correct answers. In addition, researchers indicate that the database from which the information is retrieved is incomplete. Therefore, chatbots are not able to refer to current problems, discoveries and observations in the field of science. There is also a risk of misunderstanding the question asked or the inability to create an opinion based on emotions/feelings that the chatbot does not have. Dönmez and colleagues [12] note that the lack of transparency of AI models makes it difficult for researchers to verify the accuracy and neutrality of the content provided by chatbots. [13] drew attention to the risk of excessive dependence on AI tools or limiting independent thinking. The results of the conducted research indicate that most participants accepted all information presented by chatbots as true and did not feel the need to verify their credibility.

The literature highlights the wide applications of AI in education and research, drawing attention to improving the efficiency and quality of teaching and research processes. However, as emphasised by [14–16], ethical aspects should be considered and standards for assessing the effectiveness of these technologies should be developed in order to ensure their sustainable development and implementation. The use of AI challenges teachers to constantly question the generated material and teach students to critically approach the content generated by AI. In the context of scientific work, [12] emphasise that although AI will not replace the full work of a scientist, it can significantly improve the processing and organisation of large data sets. This is because AI does not perform the scientific work for the user, but provides some guidance and support.

3 Current Study

Despite the growing integration of AI technologies in higher education, the voices of academic teachers, as key stakeholders tasked with implementing these tools in the classroom, have received limited attention in scholarly discourse. Current research predominantly examines the impact of AI on student performance [17], institutional efficiency [18], or broader ethical dilemmas [19], often overlooking the lived experiences of educators who navigate these innovations on a daily basis. This gap is particularly significant as teachers' perspectives reveal practical challenges, pedagogical adaptations, and unintended consequences that shape the effectiveness of AI implementation.

In the article, we will present the results of a study conducted among academic staff at a Polish pedagogical university. The study explores academics' attitudes towards using chatbots and large language model (LLM)-based tools in teaching and research, their existing patterns of AI use in professional work, and their perspectives on the future of this technology in teaching and research. The following research questions guide the research:

1. What patterns emerge in the use of AI-based tools by university teachers at a pedagogical university?
2. How do university teachers rate their own knowledge of, and willingness to learn about, the use of chatbots and other AI-based tools?
3. What are university teachers' attitudes towards using AI-based tools?
4. What concerns do university teachers have about using AI-based tools?
5. How do university teachers perceive the ethical issues that arise from using AI-based tools?

4 Method

4.1 Procedure

The university with which we are affiliated employs 282 people in research and teaching positions. Of these, 181 officially declare pedagogy as their research discipline. We approached these academics the intention of inviting them to participate in the study. We distributed a survey questionnaire to them in March 2025 via an anonymous link or a QR code, to use depending on the respondent's preference. Respondents were

informed that the study was voluntary and anonymous, and that they could withdraw at any time. The survey questionnaire was developed and distributed in Polish using Qualtrics software. Due to the experience of the low return rate of surveys distributed by the traditional method through the university mailing, we decided to send individual requests to academics. An unintended consequence of this action was that some participants spontaneously shared the link with other staff, not necessarily declaring pedagogy as a research discipline.

The questionnaire contained five blocks of closed questions and five open-ended questions. Two blocks of closed questions addressed the familiarity and frequency of use of selected AI chatbots and the familiarity and frequency of use of tools based on LLMs [20], with a four-point scale ranging from 1 (unfamiliar) to 4 (familiar and frequently used). One block of closed-ended questions addressed the frequency of use of selected areas of AI chatbot functionality, with the 18 author-developed items rated on a five-point scale ranging from 1 (never use) to 5 (use regularly). Additional two blocks of closed questions addressed the knowledge, attitude, concern, use of chatbots/tools based on LLMs, perceived ethics rated on a five-point Likert's scale from 1 (strongly disagree) to 5 (strongly agree) [20, 21]. Five open-ended questions developed by the authors covered topics such as experiences with chatbots and AI language tools, opinions on the integration of artificial intelligence in formal education, challenges and dilemmas of using AI in higher education.

4.2 Participants

101 participants took part in the study. We discarded records with less than 50% completion. Finally, we included data collected from 92 participants in the analyses, including 63 (68.5% of all participants) representing the discipline of pedagogy (34.8% of the targeted group for the study) and 29 participants (31.5%) representing other disciplines (psychology: 6 participants, sociology: 11 participants, philosophy: 3 participants, other: 9 participants). Specifically, the sample included 65 women (70.7%, $M_{\text{age}} = 43.1$, $SD = 9.6$, range 26–73), 25 men (27.2%, $M_{\text{age}} = 48.0$, $SD = 12.5$, range 31–77) and 2.2% of those who preferred not to provide gender information. In terms of academic degrees, the sample consisted of 26 participants with a Master's degree (28.3%), 47 participants with a Doctorate (51.1%), 16 with a Postdoctoral degree (Habilitated doctor, 17.4%) and 3 participants with a Full Professor's degree (3.3%). In terms of professional experience, 21.7% of the study sample were academics with less than five years' experience in higher education, 22.8% had five to ten years' experience, 30.4% had eleven to twenty years' experience, and 25.0% had more than twenty years' experience.

4.3 Data Analysis Strategy

This study adopted an exploratory analytical framework [22] using a mixed methods design. Quantitative data from closed-ended survey questions and qualitative insights from open-ended responses were integrated to gain a deeper understanding of the research problem [23]. Data were collected using a questionnaire originally distributed in Polish, with relevant quotes later translated into English for presentation.

Statistical analyses were performed using PS IMAGO PRO 10.0 with IBM SPSS Statistics 29.0.2. Group differences between university teachers groups in terms of affiliated discipline, academic degree or professional experience regarding chatbot usage were assessed using non-parametric Kruskal-Wallis and Mann-Whitney U tests.

Open-ended responses were analysed thematically to identify teachers' motivations, experiences and contextual factors influencing their use of AI tools. To ensure reliability, the authors independently coded emergent themes and selected representative quotes. This triangulation of qualitative and quantitative methods mitigates the inherent limitations of each approach and strengthens the validity of the findings [23].

5 Results

Participants in the study indicated that they are familiar with and use LLM-based chatbots to some extent, but ChatGPT is the tool with which they are most familiar (43.5% of participants know and use it frequently, 38.0% know but use it rarely, 1.1% do not know it, compare with [24]). They were more familiar with other AI language tools such as language translation tools, such as Google Translate, DeepL Translator (62.1% of teachers know and use them frequently, 32.2% know but use them rarely, 4.6% do not know such tools), speech-to-text transcription tools, for example automatic YouTube subtitling, TurboScribe, Note, TranscribeMe (21.8% of teachers know and use them frequently, 27.6% know but use them rarely, 18.4% do not know such tools), language learning applications, for example Duolingo, Babbel (20.7% of teachers know and use them frequently, 26.4% know but use them rarely, 14.9% do not know such applications) or online support tools to improve their writing, for example Grammarly, DeepL Write, LanguageTool (17.2% of teachers know and use them frequently, 28.7% know but use them rarely, 26.4% do not know such tools).

We were also interested in what purposes and areas teachers use chatbots for. The most common use of chatbots by academics is for language translation (22.8% of participants use that functionality regularly, 19.6% often, 23.9% from time to time). A Kruskal-Wallis tests indicated that there were no significant differences between groups in terms of affiliated discipline, academic degree or length of professional experience, $p > .159$. Of the reported uses of chatbots, only this one category was reported by more than half of respondents. Other functionalities used at least occasionally ranged from 9.7% to 33.6%. Among these, the most popular were preparation for teaching: advice on teaching methods, preparation of learning materials, etc. (4.3% of academics use this functionality regularly, 14.1% often, 15.2% from time to time), support when writing articles (4.3% of academics use this functionality regularly, 7.6% often, 17.4% from time to time), and travel: advice on what to see in a particular place, where to eat best, planning a trip, etc. (2.2% of academics use this functionality regularly, 4.3% often, 18.5% from time to time). On the other hand, the least used functionalities are those related to health and lifestyle: advice on physical activity, e.g. creating training plans (82.6% never used, 7.6% use it occasionally), advice on other health and lifestyle issues, e.g. organising time, dealing with stress, etc. (77.2% never used, 12.0% use it occasionally) and teaching support: using a chatbot for teaching, e.g. presenting new topics to students, generating exercises, quizzes or online tests (69.6% never used, 16.3% use it occasionally). Of the reported uses of chatbots, only this one category was reported by more than

half of respondents. Other functionalities used at least occasionally ranged from 16.2% (health and lifestyle: advice on physical activity, e.g. creating workout plans) to 39.2% (work and business: finding legal, tax information).

In addition to the closed questions, academic teachers could add an additional response to an open-ended question, indicating other purposes/areas for which they use chatbots. Of the 92 respondents, 33 gave no answer to this question, 39 gave a negative answer (“no”, “there are no such areas”, “I don’t know”, “I don’t use”), 20 gave a broader answer. Among the areas of use of chatbot functionality were “preparing presentations for conferences - organising material presented on slides”, “paraphrasing one’s own sentences in texts”, “using AI to improve language, stylistic errors”, “obtaining quantitative data on various topics (GDP, GINI, PISA....)”, “generating ideas, slogans, graphics”, “editing videos, graphically processing photos”, “during workshops with students”, “conducting classes on the use of digital technologies with participants/attendees of the Senior Club”, “comparing information from different sources”, “experimenting with the design of own GPTs”, “testing AI prompts for relevance and structural clarity of texts”, “email correspondence”, “creating content for social media”, “entertainment, artistic activities”, “having fun with friends”.

We then explored questions about how university teachers rate their knowledge of and experience with chatbots and other AI-based tools, as well as their attitudes towards AI training. More than half of teachers declare they know how chatbots work (13.8% strongly agree, 43.7% rather agree). A Kruskal-Wallis tests indicated that there were no significant differences between groups in terms of affiliated discipline, academic degree or length of professional experience, $p > .339$. Teachers also confirm that they have read some articles or research papers on chatbots (26.4% strongly agree, 37.6% rather agree), without difference between groups, $p > .147$. They believe that students and teachers should participate in training on using AI (56.3% strongly agree, 25.3% rather agree). There was a significant difference in attitudes towards training only between groups in reference to discipline affiliation. A Mann-Whitney U tests indicated teachers who affiliated pedagogy believed to a greater extent that training was needed ($M = 4.53$, $SD = 0.73$) than academics who affiliated other disciplines of social sciences ($M = 3.86$, $SD = 1.18$, $U = 557.0$, $p = .006$, $r = 0.29$).

More than half of academics have a positive attitude towards the use of chatbots in education (17.2% strongly agree, 40.2% rather agree), without difference between groups, $p > .331$. Only one in ten respondents believe that the use of chatbots should be banned in educational institutions (5.7% strongly agree, 6.9% rather agree, 26.4% undecided), without difference between groups, $p > .299$, but they are concerned about how chatbots will affect education in the future (32.2% strongly agree, 48.8% rather agree), $p > .472$. Respondents do not have the impression that the use of chatbots is widespread among their colleagues (10.5% strongly agree, 18.6% rather agree, 57.0% undecided). A Kruskal-Wallis test indicated that there was a significant difference between the groups in terms of length of professional experience, $H(3) = 8.84$, $p = .031$. A Mann-Whitney U test showed that academics with less than 5 years’ experience were more likely to believe that the use of chatbots was widespread among their colleagues ($M = 3.78$, $SD = 1.00$) than academics with 11–20 years’ experience ($M = 3.08$, $SD = 0.69$, $U = 136.0$, $p = .012$, $r = 0.38$) and academics with more than 20 years’ experience ($M = 3.04$,

$SD = 0.82$, $U = 119.5$, $p = .012$, $r = 0.39$). Rest of differences were not significant. Every fourth respondent is in agreement that their supervisors or university have rules or guidelines for the responsible use of chatbots (11.5% strongly agree, 12.6% rather agree), no differences between groups, $p > .411$.

Looking at the ethical issues, more than 60% of those surveyed believe that using chatbots to complete assignments and exams is cheating (37.9% strongly agree, 28.7% rather agree). Almost half of the respondents feel security concerns when using a chatbot (16.1% strongly agree, 29.9% rather agree) as well as privacy concerns (15.4% strongly agree, 31.0% rather agree). A Kruskal-Wallis test indicated that there was a significant difference between the groups in terms of academic degree, $H(3) = 8.89$, $p = .031$. A Mann-Whitney U test showed that academics with Doctor's degree ($M = 3.18$, $SD = 1.15$) were less likely to feel privacy concerns ($M = 3.18$, $SD = 1.15$) than academics with Master's degree ($M = 3.87$, $SD = 1.01$, $U = 328.0$, $p = .011$, $r = 0.31$). Rest of differences were not significant. More than 70% of those surveyed believe that chatbots can be used to manipulate information or mislead students (36.8% strongly agree, 33.3% rather agree). They also believe that chatbots could increase inequalities in education (21.8% strongly agree, 37.9% rather agree). A Kruskal-Wallis test indicated that there were no significant differences between groups.

6 Discussion

The results of the conducted research clearly indicate that the majority of respondents confirm their knowledge of chatbots, although Chat GPT is the most popular, which is also confirmed by the results of other studies [3, 25]. The findings indicate a moderately high self-assessment of academic staff's knowledge and experience regarding the use of chatbots and other AI-based tools. More than half of the respondents report understanding how chatbots function, which may reflect the growing presence of AI technologies within the pedagogical university environment, as well as the increasing adoption of such tools by teaching personnel. However, it is important to note that these self-reported declarations do not necessarily equate to actual technological-pedagogical proficiency. Instead, they may reflect a general familiarity with the concept of AI operation rather than an in-depth understanding of its limitations and pedagogical potential.

The absence of statistically significant differences between the analysed groups suggests that knowledge of basic AI functions is relatively universal and not strongly determined by traditional demographic or professional variables. At the same time, it is noteworthy that there is a high level of willingness among academic staff to acknowledge the need for training in AI use in education, with as many as 81.6% of respondents expressing support for this idea. This result can be interpreted as an indication of growing awareness that effective and ethical use of AI requires more than intuitive engagement; it also demands support in developing digital competencies, pedagogical reflection, and familiarity with legal and ethical frameworks.

The most significant finding of the study is that teachers specialising in pedagogy are more likely to recognise the need for training than their colleagues from other social science disciplines. This difference was statistically significant and may be attributed to the greater emphasis within the field of pedagogy on the development of didactic competences, critical engagement with tools that support teaching and learning processes,

and professional reflexivity. Educators in this field may also have more frequent and direct contact with topics related to educational transformation and the technological implications of working with learners, which reinforces their belief in the necessity of being adequately prepared for the informed use of AI.

Using tools based on artificial intelligence relieves teachers of routine tasks and provides great support in creating educational materials used during teaching classes. Additionally, chatbots allow teachers to adapt teaching methods to the individual needs of each student, which affects not only the efficiency of teachers' work but also more effective and more engaged learning, which is also emphasised by other researchers [7]. Academic teachers also use AI tools for development. They most often use chatbots for language translation and some use AI to help prepare scientific articles or conference presentations. AI technology is also used for text editing. Chatbots can paraphrase sentences in texts and correct language or stylistic errors. Both chatbots and language tools like Grammarly are becoming increasingly popular aids in their professional work, mainly due to changing language trends. A large proportion of national scientific journals publish articles in English, and more and more scientists are seeking to publish texts in foreign journals, so support in developing language skills is becoming a regular feature of researchers' work.

Teachers have a positive attitude towards the use of chatbots. To a greater extent, this is the result of bottom-up, individual initiatives to use AI in teaching or research, as only a quarter of staff believe that their university has introduced rules and guidelines for the responsible use of chatbots and LLM-based tools in education. As other studies have shown, this problem is still very much alive [20, 26]. In the context of shaping future generations, it is crucial to find a balance between using the advantages of artificial intelligence and maintaining the important aspects of traditional education, which shape social skills, empathy and the ability to think critically. Responsible and conscious implementation of AI technology in the education system will have a decisive impact on the development and preparation of young people for future challenges.

An important aspect is ethics when using AI technology. It is true that few teachers believe that the use of chatbots in educational institutions should be banned, if only because it misleads students or is used to manipulate information. The use of chatbots in preparing tasks and writing exams by students raises great doubts. More than half of the respondents indicated that these are unfair practices. There is also a concern that too much reliance on technology may limit independence and the ability to think critically, as indicated by other research results [9]. To prevent this, both teachers and students should participate in training to deepen their knowledge and expand their skills in using AI technologies. Based on the obtained research results, a significant difference should be pointed out, academics affiliated with pedagogy see a greater need for AI training than academics affiliated with other disciplines.

Another aspect of the use of technology by academics is the sense of security. When using chatbots, teachers feel concerns about security and the protection of their privacy. However, as the results of the conducted research indicate, teachers with a master's degree had significantly more concerns about security than those with a doctorate.

7 Conclusion

This study examines how academics at a Polish pedagogical university use chatbots and AI-based language tools in their professional work training future teachers. The findings reveal the interplay between familiarity, usability and concerns that shape academic engagement with AI-based tools. Looking at patterns of adoption and use, academics show high familiarity with available chatbots, particularly ChatGPT, while use of other AI language tools remains moderate. Language translation emerges as the dominant application, used by more than half of respondents, while pedagogical support such as lesson preparation and material development is moderate. On the other hand, direct classroom integration (like using chatbots for teaching) and health/lifestyle advice are the least used, highlighting the preference for chatbots in supporting academic tasks rather than in core educational or personal domains.

In terms of perceptions, knowledge and training needs for AI tools, despite more than half of respondents understanding the mechanics of chatbots and engaging with relevant literature, there is a strong consensus among education stakeholders on the need for AI training. In particular, disciplinary differences are evident, with academics affiliated with pedagogy placing more emphasis on training than their counterparts from other social sciences, highlighting a discipline-specific willingness to integrate AI. While 50% view chatbots in education positively, 10% support institutional bans, reflecting an ambivalence rooted in concerns about academic integrity and future implications for education.

In response to the research question on ethical and institutional issues, our study shows that ethical concerns are prominent, approximately 60% perceive chatbot-assisted tasks as cheating, and 70% warn of the risk of misinformation and deepening of educational inequalities. Security and privacy concerns affect almost half of respondents, with academics with master's degrees expressing greater privacy concerns than those with doctoral degrees. This discrepancy may be related to experience or institutional roles. Worryingly, only around 25% of respondents are aware of institutional policies on the responsible use of technology, suggesting a critical gap in university policy that could hinder the ethical adoption of AI in the workplace and the dissemination of AI knowledge to colleagues and students.

Early career researchers (less than five years) perceive a greater use of AI among colleagues compared to older colleagues, suggesting a generational shift in AI adoption. This discrepancy highlights the need for mentoring and sharing of knowledge and experience to harmonise the adoption of AI tools across different levels of experience.

Although this study provides valuable insights into academics' perceptions and use of chatbots in education, it has several limitations. The study was conducted using a self-administered questionnaire, which tends to capture prevailing patterns, but does not explore in depth the sources or arguments behind the attitudes expressed by academics. Future research should therefore be extended to include in-depth qualitative interviews that help to clarify these issues and better understand the concerns behind the limited confidence in the use of AI in the education of future teachers. In terms of the research sample, the study focused on academics at a selected public teacher training university, so the findings cannot be generalised. It is advisable to continue the study in other Polish

pedagogical universities, both in smaller cities and non-public ones, in order to verify the replicability of the obtained results.

Despite the limitations highlighted, this study makes an important contribution to filling a gap in the literature by providing the perspective of those at the forefront of implementing artificial intelligence into the educational process of current and future generations. The uniqueness of the study lies in the selection of the research sample, staff members of pedagogical universities who have a direct impact on the education of future teachers.

Authors' Contribution. Conceptualisation: E.D.; Methodology: E.D.; Formal analysis: E.D.; Investigation: E.D., E.M.-K., and J.W.-M.; Data curation: E.D.; Writing—original draft preparation: E.D., E.M.-K., and J.W.-M.; Review and editing: E.D., E.M.-K., and J.W.-M.; Supervision: E.D. All authors have read and agreed to the published version of the manuscript.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

- Hayes, A.: AI for digital transformation: strategies for business resilience and growth. *Int. J. Uniq. New Updates* **4**(1), 16–23 (2022)
- Karataş, F., Eriçok, B., Tanrikulu, L.: Reshaping curriculum adaptation in the age of artificial intelligence: Mapping teachers' AI-driven curriculum adaptation patterns. *Br. Edu. Res. J.* **51**(1), 154–180 (2025)
- Lim, W.M., Gunasekara, A., Pallant, J.L., Pallant, J.I., Pechenkina, E.: Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The Int. J. Manage. Edu.* **21**(2), 100790 (2023)
- Firth, D.R., Triche, J.: Generative AI in practice: a teaching case in the introduction to management information systems class. *Info. Sys. Edu. J.* **22**(4), 29–47 (2024). <https://doi.org/10.62273/LDVL8354>
- Lund, B.D., Wang, T.: Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News* **40**(3), 26–29 (2023)
- Kuleto, V., et al.: Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions. *Sustainability* **13**(18), 10424 (2021)
- Rasul, T., et al.: The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *J. Appl. Learn. Teach.* **6**(1), 41–56 (2023)
- Sullivan, M., Kelly, A., McLaughlan, P.: ChatGPT in higher education: Considerations for academic integrity and student learning. *J. Appl. Learn. Teach.* **6**(1), 1–10 (2023). <https://doi.org/10.37074/jalt.2023.6.1.17>
- Bin-Nashwan, S.A., Sadallah, M., Bouteraa, M.: Use of ChatGPT in academia: academic integrity hangs in the balance. *Technol. Soc.* **75**, 102370 (2023)
- Karamanhoğlu, A.U.: AI in academic research: advances, opportunities, and challenges. In: Tripathi, S., Rosak-Szyrocka, J. (eds.) *Impact of Artificial Intelligence on Society*, pp. 81–97 (2024)
- Humble, N., et al.: Cheaters or AI-enhanced learners: Consequences of ChatGPT for programming education. *Electr. J. e-Learning* **22**(2), 6–29 (2024). <https://doi.org/10.34190/ejel.21.5.3154>

12. Dönmez, İ., Idin, S., Gülen, S.: Conducting academic research with the AI interface ChatGPT: Challenges and Opportunities. *J. STEAM Edu.* **6**(2), 101–118 (2023). <https://doi.org/10.55290/steam.1263404>
13. Dilekli, Y., Boyraz, S.: From 'Can AI think?' to 'Can AI help thinking deeper?': Is use of ChatGPT in higher education a tool of transformation or fraud? *Int. J. Modern Edu. Stud.* **8**(1), 49–71 (2024). <https://doi.org/10.51383/ijonmes.2024.316>
14. Kumar, R.: Beyond approach: Navigating usage, detection, and future pathways of AI in education. *A J. Edu. Res. Pract.* **33**(3), 22–29
15. Guilherme, A.: AI and education: The importance of teacher and student relations. *AI & Soc.* **34**(1), 47–54 (2019). <https://doi.org/10.1007/s00146-017-0693-8>
16. Cardona, M.A., Rodríguez, R.J., Ishmael, K.: Artificial intelligence and the future of teaching and learning: Insights and recommendations. *Education Next* (2023)
17. García-Martínez, I., Fernández-Batanero, J.M., Fernández-Cerero, J., León, S.P.: Analysing the impact of artificial intelligence and computational sciences on student performance: Systematic review and meta-analysis. *J. New Approach. Educat. Res.* **12**(1), 171–197 (2023)
18. Srivastava, P., Hassija, T., Goyal, A.P.: Unleashing the Potential of Artificial Intelligence in the Education Sector for Institutional Efficiency. In: *Transforming Management Using Artificial Intelligence Techniques*, pp. 11–22. CRC Press, Boca Raton (2020)
19. Sain, Z.H., Aziz, A.L., Lawal, U.S.: Ethics of AI Integration in Higher Education: Exploring Moral Dimensions. *J. Asian Islamic Educat. Manage.* **2**(1), 1–10 (2024)
20. Malmström, H., Stöhr, C., Ou, A.W.: Chatbots and other AI for learning: A survey of use and views among university students in Sweden. *Chalmers Studies in Communication and Learning in Higher Education* **2023**, 1 (2023). <https://doi.org/10.17196/cls.cslche/2023/01>
21. Acosta-Enriquez, B.G., Arbulú Ballesteros, M.A., Arbulu Perez Vargas, C.G., et al.: Knowledge, attitudes, and perceived Ethics regarding the use of ChatGPT among generation Z university students. *Int. J. Educat. Integ.* **20**(1), 10 (2024)
22. Hochbein, C., Smeaton, K.S.: An exploratory analysis of the prevalence of quantitative research concepts in journal articles. *Int. J. Edu. Policy and Leaders.* **13**(11) (2018)
23. Lund, T.: Combining qualitative and quantitative approaches: Some arguments for mixed methods research. *Scand. J. Educ. Res.* **56**(2), 155–165 (2012)
24. Duda, E., Młynarczyk-Karabin, E., Więckiewicz-Modrzewska, J.: University Teachers towards ChatGPT. Challenges and Opportunities. *Proceedings of the Ethicomp 2025. Universidade Autónoma de Lisboa, Portugal* (2025)
25. Stahl, B.C.: *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*. Springer, Cham (2021)
26. Neumann, M., Rauschenberger, M., Schön, E.M.: We need to talk about ChatGPT: The future of AI and higher education. In: *2023 IEEE/ACM 5th International Workshop on Software Engineering Education for the Next Generation (SEENG)*, pp. 29–32. IEEE (2023)



Ethical Issues in the Use of Generative AI Chatbots for Therapeutic Purposes

Anne Gerdes^(✉) 

University of Southern Denmark, Universitetsparken 1, 6000 Kolding, Denmark
gerdes@sdu.dk

Abstract. This paper discusses the pros and cons related to the digital transformation of mental health, particularly highlighting the therapeutic domain; against this setting, the technical underpinnings of generative AI chatbots are outlined alongside a discussion of the broader ethical and epistemic challenges related to the integration of LLMs in the medical domain—a domain that is more advanced in experimenting with and deploying LLMs than the field of mental health. While focused on mental health, the paper draws on these broader discussions to highlight shared concerns, complemented by insights specific to mental health. Against this backdrop, the paper analyzes AI-based therapy and generative AI companion chatbots, assessing their potential benefits and associated risks, including considerations of relevant European legal frameworks. Following up, the paper argues that even if we succeed in escaping the challenges at the intersection of ethics, episteme, and technology, generative AI chatbots are ill-suited for therapeutic purposes. Playing with the idea of replacing human relationships with “stochastic parrots” risks reducing therapy to algorithmic risk stratification. Such a shift threatens to diminish the potential for cultivating authentic, trust-based therapeutic relationships.

Keywords: Generative AI chatbots · Authenticity · Therapy · LLMs · European legislative frameworks

1 Introduction

This paper starts by outlining the pros and cons associated with the overall digital transformation of mental health, particularly focusing on the therapeutic domain (Sect. 2). In section three, the paper outlines the technical foundation for generative AI chatbots by examining the concept of foundation models and by addressing the broader ethical and epistemic challenges related to the integration of LLMs in the medical domain, as this field has seen more extensive experimentation and deployment of LLMs than the field of mental health. While the paper focuses on mental health, these broader discussions are introduced because they highlight concerns that are also at stake in the domain of mental health. In addition to addressing these general challenges, the analysis draws on emerging literature on LLMs in mental health and domain-specific insights. Against this setting, the paper turns to AI therapy bots and generative AI companion

bots and explores their potential benefits and risks (Sect. 3), including attention to relevant legislation (Sect. 4). Clearly, we should not underestimate the efforts it will take to safely develop and deploy such systems. Still, the paper asks the question of whether we should apply generative AI chatbots for therapeutic purposes, given that we succeed in escaping the challenges at the intersection of ethics, episteme, and technology. To answer this question, the paper discusses the role of authenticity in human-generative AI chatbot interactions (Sect. 5), concluding that (Sect. 6) generative AI chatbots are ill-suited for therapeutic purposes, as authenticity constitutes a foundational element in therapy. Owing to their lack of lived experience, such systems cannot replicate therapeutic relationships. Applying generative AI chatbots for therapeutic purposes risks turning therapy into algorithmic risk stratification, thereby undermining human dignity and the opportunity for building trustful, authentic relationships.

2 The Digital Transformation of the Therapeutic Field

Mental health issues are growing significantly worldwide [30], and ensuring easy and equitable access to treatment, especially in remote areas, is a challenge. Therefore, digital solutions are seen as an essential tool for mental health promotion. AI-based online therapy, possibly with psychological support, can reduce waiting times and alleviate mild forms of depression and anxiety [8]. Similarly, AI tools can streamline screenings and aid in diagnostics, enabling earlier intervention in a patient's illness trajectory. Online therapy can also reduce stigma, as these services are perceived as more anonymous and thus less intimidating to engage with. Furthermore, younger generations expect digital treatment options to be accessible at their convenience. Some even refer to a "mute generation" that prefers text-based communication over verbal interaction [6].

However, digital solutions also bring several ethical challenges. Security breaches and data leaks can have significant consequences for individuals. In addition, privacy violations are not solely related to data leaks; they can also arise when users feel unable to understand the terms they consent to. Mental health apps may seem trustworthy because people often associate them with the credibility of healthcare services. These apps appear more serious and professional than general apps. As such, they can mislead users into believing they offer a confidential therapeutic space, even when their data is sold for other purposes [16].

Another concern is whether digital solutions might eventually replace the authentic interaction between therapists and their patients. In remote areas, online services can improve access to treatment and promote mental health equity. Yet, such initiatives could lead to digital solutions becoming the only option in these areas due to a lack of incentives to maintain alternative services. Finally, there is widespread optimism about AI, which downplays the fact that developing AI solutions is labor-intensive and requires the involvement of domain experts with deep knowledge of therapeutic and clinical practice.

To set the scene before turning to generative AI chatbots, the paper outlines their technical foundation by examining the concept of foundation models and reflects on the broader ethical and epistemic challenges introduced by LLMs in the healthcare sector. While the paper focuses on mental health, these broader discussions are included because

they highlight concerns also at stake in the domain of mental health, which lags behind the domain of medicine in experimenting with and deploying LLMs [15]. The analysis further draws on emerging literature on LLMs in mental health and domain-specific insights.

3 Foundation Models - the Promises and Perils of LLMs in Healthcare

In recent years, machine learning has undergone a significant transformation fueled by the rise of powerful foundation models (FMs), such as large language models (LLMs), e.g., GPT or Llama, and diffusion models, e.g., Stable Diffusion XL or Midjourney for image generation. Here, we concentrate on FMs for text generation. LLMs are built to perform string prediction tasks, trained on extensive corpora of unannotated text using unsupervised reinforcement learning with human feedback. This training enables deep neural networks to estimate the probability of individual tokens – that is, discrete and meaningful linguistic units – based on their surrounding context. The foundational concept of statistical language modeling is far from new but originated in the mid-20th century. For instance, in 1948, Shannon outlined the stochastic process governing a “physical system” or “a mathematical model” in carrying out string prediction tasks [20]. The message will be generated “symbol by symbol”, and the model “will choose successive symbols according to certain probabilities depending, in general, on preceding choices as well as the particular symbols in question” [20].

Despite these early insights, LLMs didn’t gain momentum until the recent advancement in data availability, computational capacity, and storage infrastructure. A major breakthrough is credited Vaswani et al. [26], who introduced the transformer architecture based on attention mechanisms. This design enables models to process and learn from large-scale, unlabeled data by capturing long-range dependencies across sequences and dynamically weighing the relationships between tokens within a given context.

Training a base model is labor-consuming and has substantial economic and environmental costs [4, 21]. FMs can also be trained by applying smaller domain-specific data sets from scratch, such as PubMEDBERT. However, due to the scarcity of medical data, it can be challenging to obtain models that generalize well [31]. Once the base model has learned the statistical patterns of languages, it can be fine-tuned for various downstream tasks by feeding it curated domain-specific data sets and applying human feedback by domain experts in the evaluation of the model’s output. Still, even without domain-specific training, models have demonstrated outstanding capabilities for zero-shot or few-shot predictions when facing totally new tasks, meaning that they can produce correct output without further re-training and finetuning. For example, “without medical-specific training,” GPT4 passed the United States Medical Licensing Examination in 2023 and can pass examinations in specialist areas and diagnosing challenging cases [23].

There are many promising use scenarios for FMs to tackle the ever-growing challenges in the healthcare sector. For instance, LLMs can support patient-healthcare interactions by enhancing patient communication tailored to patients’ unique needs [19]. As such, health equity can be promoted by personalizing information that “meets a patient’s

literacy level and preferred language” [18]. On a global level, Botelho et al. [7] suggest that LLMs have the potential to assist healthcare professionals in low and middle-income countries.

LLMs can be helpful for diagnostic support, extracting information from health records, summarizing medical evidence, ranking treatment options [14], and drafting responses in e-consultations, contributing to decreasing the workload for healthcare professionals [2]. Furthermore, Lawrence et al. [15] suggest that LLMs “trained in EBP [evidence-based practice] could provide in-the-moment support to therapists providing telehealth.” Moreover, in medical education, LLMs may facilitate personalized learning [19].

FMs have the potential to transform healthcare, yet due to their complexity, malleability, and scale of use and reuse, they exacerbate well-known ethical and epistemic issues in machine learning. A base model trained on data from the internet preserves biases inherent in the training data, and a subsequently fine-tuned model might inherit such semantic biases, obscuring equity in health by generalizing in a manner that causes representational harm to marginalized groups. Within mental health, Lawrence et al. [15] emphasize existing health inequities. Mental health issues are experienced as stigmatized, and people of low socioeconomic status are more at risk for developing mental health disorders. Furthermore, we witness unequal access to mental health care. Lawrence et al. [15] note that much of the available information concerning mental health, as well as the discourse surrounding mental health, is marked by inaccuracies and stigma, while the existing research literature predominantly reflects the perspectives of individuals who are White, highly educated, economically privileged, and English-speaking. As a result, LLMs might amplify health inequity.

Moreover, FMs can generate factual inconsistent information. Thus, the faithfulness problem encapsulates the epistemic challenging flip side of FMs’ potential to impact medical knowledge and health communication positively. Consequently, the fact that FM always appears overly convincing despite being sometimes wrong may lead to harm due to misinformation. Factual inconsistencies may be caused by scarcity of medical training data or data discrepancies between ground truth outputs used during training and subsequent updated or revised medical knowledge [31].

The fact that we cannot rely on FM output may negatively influence informed decision-making and, furthermore, hamper patient-centered decision-making if FM-generated outputs are untrustworthy [5]. Finally, the ethically responsible introduction of LLMs in healthcare requires open FMs. When LLMs are marketed by corporations inclined toward disruptive business practices, implementing FMs in the healthcare sector or any other high-stake domain is irresponsible – “it is hard to see how LLMs that are developed and controlled behind closed corporate doors could be broadly adapted in healthcare without undermining the accountability and transparency of both medical research and medical care” [23].

Against this setting, the following section zooms in on generative AI chatbots in discussing their potentialities in the field of therapy, followed by an introduction to the legislative framework surrounding these systems.

4 Generative AI Chatbots - Promises, Risks, and the Role of Regulation

Current AI therapy bots, such as Woebot, use advanced, rule-based scripts designed in collaboration with cognitive behavioral therapy experts. These chatbots can somewhat reduce stress and anxiety symptoms [15] and assist with fighting alcohol abuse [17]. Users often perceive these therapeutic bots as empathetic despite their conversational style being less fluent than that of generative AI chatbots. So, it might improve therapy if LLMs enter the scene, and already now, many generative AI chatbots are used for therapy and self-diagnosis.

For example, the AI companion Replika, based on the GPT model and cognitive therapy principles, is marketed as a lifelong friend. Although the developers emphasize that Replika is not a therapy substitute, many users treat it as such. While users may experience emotional support from generative AI companion chatbots, interactions can also foster dependency, worsen mental health conditions, or have severe consequences. For instance, the “Character. AI” platform allows users to converse with fictional characters. Allegedly, this led to a 14-year-old boy being manipulated into suicide, prompting a lawsuit against Character.AI and its founders, who now work at Google [11]. Google has disclaimed responsibility, noting that they are unrelated to Character.AI.

The market for low-quality mental health apps and generative companion chatbots is likely to expand as developing such services becomes more effortless. However, EU legislation aims to mitigate wrongdoing by safeguarding personal data through the General Data Protection Regulation (GDPR) [13], by requiring providers of intermediary services and online platforms to comply with the Digital Service Act (DSA) [10], and by introducing a risk-based approach under the AI Act [1].

Consequently, from August 2026, AI systems within the healthcare sector are classified as high-risk by default (Article 6, Annex III)¹. The development and deployment of high-risk AI systems are described in Title III, Articles 6–51, and require proactive management and handling of ethical and safety issues, including, among other things, AI literacy requirements (Article 4), risk management (Article 9), data and data governance (Article 10), technical documentation (Article 11), record keeping (Article 12), transparency and provision of information to users (Article 13), human oversight (Article 14), accuracy, robustness, and cybersecurity (Article 15), and responsibility and impact assessment of the AI system’s effect on fundamental rights (Articles 26 and 27). As such, AI systems dedicated to clinical decision-making or therapeutic purposes are expected to fall in the high-risk category (Article 6, Annex III).

The AI Act may thus help stem the spread of low-quality AI products marketed as therapeutic solutions, as such systems are expected to be classified as high-risk AI. AI companions intended for entertainment purposes are, by default, classified as AI systems posing a *limited risk*, for which the primary requirement is that users must be informed they are interacting with a machine.

¹ Although an AI system is used within a high-risk domain, certain specific applications are not considered high-risk because their use does not pose a significant risk. This applies, for example, when the AI systems in the healthcare sector are used to carry out a narrowly defined procedural task, such as classifying incoming documents into categories (AIA, Article 6(3)).

However, the Act also defines a category of *prohibited practices*—namely, AI systems that pose an *unacceptable risk* (AI Act, Title II), [1]. These include AI systems that employ subliminal techniques “beyond a person’s consciousness in order to materially distort a person’s behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm” (Article 5(1)(a)). Moreover, prohibited practices also include AI systems capable of manipulating vulnerable groups, such as children or individuals with disabilities, in ways that may result in physical or psychological harm (Article 5(1)(b)). In this context, a generative AI companion chatbot will presumably be deemed an *unacceptable risk* if it begins to behave in a manner resembling that of Character.AI [12].

Additionally, these systems might also fall under the scope of the Digital Service Act [10] if they are integrated into or hosted by a platform, such as a very large online platform (VLOP), which will have to live up to transparency requirements and moderator obligations and shall have to introduce safeguards aimed at reducing systemic risks and carrying out risk assessments to evaluate potential psychological impacts, addiction potentialities, and data protection concerns.

In summary, the European legislative frameworks can presumably promote the responsible development of generative AI chatbots.

Still, regulation alone is insufficient, as we continue to witness online moral wrongdoing despite legislative efforts. Therefore, we must prioritize facilitating users’ development of digital literacy skills, allowing them to engage critically with technology. However, this topic reaches beyond the scope of this paper, although the author acknowledges the pressing need to enhance digital literacy in our societies.

Summing up, in sections three and four, the promises and perils of LLMs in mental healthcare have been laid out as a backdrop for considering their future integration into mental health, particularly zooming in on generative AI chatbots to answer the question of whether we should embrace such systems for therapeutic purposes. Following Lawrence et al. [15], we might do so, given that the above-mentioned challenges have been solved. They argue that when LLMs are rigorously tested and aligned with evidence-based practices (EBP), and “the goals of people with lived experience” as well as professionals with expertise in mental health care, they hold significant promise for expanding access to evidence-based mental health information and services. According to the authors:

“Investing in developing, testing, and deploying mental health LLMs responsibly has the potential to finally reverse rising global mental health rates and to improve the mental health of the millions of people in need of mental health support.” [15]

5 Generative AI Chatbots – the Question of Authenticity

Assuming that, in the future, we will succeed in escaping the abovementioned challenges at the intersection of ethics, episteme, and technology. Suppose that we become able to develop reliable and ethically aligned generative AI chatbots for therapeutic purposes that live up to legislative frameworks and fulfill rigorous safety, transparency, and accountability standards. And suppose that people have a high level of AI literacy. In that case, should we then follow the advice of Lawrence et al. [15] and apply generative AI chatbots for therapeutic purposes?

Let us revisit the 1960s when computer scientist Joseph Weizenbaum developed the ELIZA program [28]. This early chatbot included a version called DOCTOR, based on the principles of Carl Rogers' psychotherapy. Client-centered therapy is relatively straightforward to simulate because the therapist repeats the client's own words.

Weizenbaum was profoundly unsettled by the attention ELIZA earned. He identified three particularly disturbing reactions to ELIZA. First, psychiatrists expressed admiration for the potential of automated therapy, with some claiming that computers could eventually replace and surpass therapists. They even described the therapeutic task in computer terms as "information processing." Second, people anthropomorphized ELIZA excessively by having deep conversations with the simple program. Third, he was astonished that people believed ELIZA demonstrated genuine language comprehension. He feared technological developments could lead to a loss of authenticity, dehumanization, and a mechanical understanding of humanity [29].

The notion of human-centeredness is critical in the intersection of authenticity [22] and generative AI chatbots. Some studies tentatively suggest that patients find language models more empathetic than interactions with doctors, e.g., [24]. From a utilitarian perspective, artificial empathy can be justified if it achieves good clinical outcomes and benefits most people. However, the ends do not always justify the means. It may be ethically problematic to rely on artificial empathy from what has been called a "stochastic parrot" [4].

Taylor highlights the dialogical nature of human existence as a precondition for the ideal of authenticity, emphasizing that authenticity requires "openness to horizons of significance" [22]. When it comes to authenticity, I must, of course, first and foremost be in contact with myself in order to be true to myself. However, this does not imply a subjectivist conception of self-realization as a solipsistic engagement, reducible to an isolated dialogue between myself and a generative AI chatbot. According to Taylor, authenticity must be understood in light of our dialogical condition, in which our relationships with others place demands upon us – thus, being in contact with myself requires "self-definition in dialogue" [22]. Taylor argues that our authentic self is shaped through dialogical interactions with significant others in our environment. These significant others cannot simply be reduced to automated therapists supporting individualized self-realization.

On the other hand, Bassett suggests that the Rogerian-inspired structure underlying the DOCTOR version of Eliza can also be understood as a tool for self-actualization [3]. This implies that Weizenbaum's dichotomy between the computer and human autonomy is overly simplistic, since, according to Bassett, a socio-technical relationship can create a therapeutic space for "self-realized autonomy." Similarly, Verbeek [27] conceptualizes technologies as mediating artifacts that contribute to shaping our understanding of the world and our actions within it. Although technological artifacts do not possess human intentions, Verbeek operates with the notion of distributed or hybrid intentionalities, whereby technology can guide (Latin: *intentio*) and co-constitute human action and experience.

However, these considerations do not imply that Weizenbaum was mistaken in his overarching claim that dependence on computers can lead to the loss of authentic human interactions and a devaluation of human judgment. It is, therefore, important to consider

not only the manner in which we engage in dialogue but also the entities with whom we converse. In a similar vein, Turkle warns against the rise of a “robotic moment,” in which individuals may begin to favor the safety and predictability of interactions with AI companions over the complexities and demands of human relationships [25].

With their realistic communication capabilities, generative AI chatbots invite us to share our inner lives with a stochastic parrot [4]. Nevertheless, the authentic connection between therapist and client remains a foundational element of therapeutic practice—something generative AI chatbots, lacking lived experience, are inherently unable to replicate.

6 Conclusion

This paper has outlined the promises and perils of LLMs in mental health as well as introduced the overall legislative European frameworks to investigate whether we should apply generative AI chatbots for therapeutic purposes. We should not underestimate the efforts it will take to safely deploy generative AI chatbots in therapy. Currently, while LLMs are promising, they are unreliable and can provide harmful therapeutic advice; they might exacerbate disparities, health inequities, biases, and fuel discriminatory practices. There is still a long way to go before LLMs can be responsibly used in therapeutic contexts. Yet, this paper has also asked the question of whether we should deploy generative AI chatbots for therapeutic purposes if we become capable of developing LLMs that live up to technical, ethical, and legal standards. Clearly, doing so would help alleviate the burdens within mental health care, and the growing global mental health needs could be met by responsible development and deployment of LLMs in mental health.

However, even though we play with the idea that all challenges evaporate, we should not forget that we also play with the idea that human relationships can be replaced by stochastic parrots. We risk turning therapy into algorithmic risk stratification, thereby undermining the opportunity for building trustful, authentic relationships.

Disclosure of Interests. The author has no competing interests to declare relevant to this article’s content.

References

1. AI Act - Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act (Text with EEA relevance) (2024). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
2. Ayers, J.W., et al.: Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine* **183**(6), 589–596 (2023). <https://doi.org/10.1001/jamainternmed.2023.1838>
3. Basset, C.: The computational therapeutic: exploring Weizenbaum’s ELIZA as a history of the present. *AI & Soc.* **34**(4), 803–812 (2019)

4. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT'21: Proceedings of the 2021 ACM CONFERENCE on FAIRNESS, ACCOUNTABILITY, and TRANSPARENCY, virtual event Canada (2021)
5. Bjerring, J.C., Busch, J.: Artificial Intelligence and Patient-Centered Decision-Making. *Philosophy & Technology* **34**(2), 349–371 (2021). <https://doi.org/10.1007/s13347-019-00391-6>
6. Bond, R.R., et al.: Digital transformation of mental health services. *NpJ Mental Health Research* **2**(1) (2023). <https://doi.org/10.1038/s44184-023-00033-y>
7. Botelho, F., Tshimula, J.M., Poenaru, D.: Leveraging ChatGPT to democratize and decolonize global surgery: large language models for small healthcare budgets. *World J. Surg.* **47**(11), 2626–2627 (2023)
8. Center for Digital Psykiatri.: <https://psykiatrienisyddanmark.dk/afdelinger/specialiserede-afdelinger-og-centre/center-for-digital-psykiatri>. last accessed 17 April 2025
9. Chaturvedi, R., Verma, S., Das, R., Dwivedi, Y.K.: Social companionship with artificial intelligence: recent trends and future avenues. *Technol. Forecast. Soc. Chang.* **193**, 122634 (2023)
10. DSA - Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Service Act) (text with EEA relevance) PE/30/2022/REV/1 (2022). <https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng>
11. Duffy, C.: CNN: 'There are no guardrails.' This mom believes an AI chatbot is responsible for her son's suicide (2025). <https://edition.cnn.com/2024/10/30/tech/teen-suicide-character-ai-lawsuit/index.html>. Accessed 17 May 2025.
12. European Commission: Guidelines on prohibited artificial intelligence (AI) practices defined in the AI Act. Publications Office of the European Union (2024). <https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act>. Accessed 17 May 2025
13. GDPR - Regulation (EU) 2016/679 of the European Parliament of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) (2016). <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>
14. Grote, T., Berens, P.: Uncertainty, Evidence, and the Integration of Machine Learning into Medical Practice. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* **48**(1), 84–97 (2023). <https://doi.org/10.1093/jmp/jhac034>
15. Lawrence, H.R., Schneider, R.A., Rubin, S.B., Matarić, M.J., McDuff, D.J., Jones Bell, M.: The opportunities and risks of large language models in mental health. *JMIR Ment Health* **11**, e59479 (2024). <https://doi.org/10.2196/59479>
16. Martinez-Martin, N., Kreitmair, K.: Ethical issues for direct-to-consumer digital psychotherapy apps: addressing accountability, data protection, and consent. *JMIR Ment Health* **5**(2), e32 (2018)
17. Prochaska, J.J., et al.: A therapeutic relational agent for reducing problematic substance use (woebot): development and usability study [original paper]. *J. Med. Internet Res.* **23**(3), e24850 (2021)
18. Rodriguez, J.A., Alsentzer, E., Bates, D.W.: Leveraging large language models to foster equity in healthcare. *J. American Medi. Info. Ass.* (2024). <https://doi.org/10.1093/jamia/ocae055>
19. Sallam, M.: ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* **11**(6) (2023). <https://doi.org/10.3390/healthcare11060887>

20. Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948)
21. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP. the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy (July 2019). [arXiv:1906.02243v1](https://arxiv.org/abs/1906.02243v1)
22. Taylor, C.: *The Ethics of Authenticity*. Harvard University Press, Massachusetts (1991)
23. Toma, A., Senkaiahliyan, S., Lawler, P.R., Rubin, B., Wang, B.: To safely deploy generative AI in healthcare models must be open source. *Nature (London)* **624**(7990), 36–38 (2023). <https://doi.org/10.1038/d41586-023-03803-y>
24. Tu, T., et al.: Towards Conversational Diagnostic AI (2024). [http://arxiv.org/abs/2401.05654](https://arxiv.org/abs/2401.05654)
25. Turkle, S.: *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books, NY (2011)
26. Vaswani, A., et al.: Attention is all you need (2017). <https://doi.org/10.48550/arXiv.1706.03762>
27. Verbeek, P.P.: *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press (2011)
28. Weizenbaum, J.: ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**(1), 36–45 (1966)
29. Weizenbaum, J.: *Computer power and human reason: From judgment to calculation*. Freeman San Francisco (1976)
30. WHO: World mental health report: transforming mental health for all. World Health Organization (2022). URL: <https://www.who.int/publications/i/item/9789240049338>. Last accessed 17 March 2025
31. Xie, Q., et al.: Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond (2023). <https://doi.org/10.1101/2023.04.18.23288752>



A Case Against the Feasibility of AI Consciousness (AIC)

Bo Kampmann Walther¹  and Anne Gerdes²  

¹ University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark

² University of Southern Denmark, Universitetsparken 1, 6000 Kolding, Denmark
gerdes@sdu.dk

Abstract. This paper argues against the feasibility of AI Consciousness (AIC), discussing recent claims that large language models and advanced AI architectures could one day possess phenomenal consciousness. While proponents draw on computational functionalism and models like Global Workspace Theory to support the idea of AIC, we challenge this view by highlighting the ontological and phenomenological conditions of consciousness. We argue that consciousness, as viewed through classical phenomenology and critical accounts of embodiment and embeddedness, is a transcendental precondition for experience. Thus, consciousness cannot emerge in silicon-based systems because it is irreducibly lived, embodied, and first-personal.

Keywords: Consciousness · Large Language Models · Embodiment · Social Embeddedness

1 Introduction

In the movie *Notting Hill*, friends of William Thacker (Hugh Grant) organize a series of dates to help him recover from the heartbreak of losing the famous actress Anna Scott (Julia Roberts). One of the dates is a fruitarian who declines the dinner dish, explaining, “We believe that fruits and vegetables have feelings, so we think cooking is cruel. We only eat what has fallen from a tree or bush and that is, in fact, dead already.” William responds, “These carrots?” “Have been murdered? –Yes.”

We will not discuss fruitarianism or fruitarian ideologies here. Instead, we argue that consciousness doesn’t extend beyond animals. Hence, like carrots, AI does not – and cannot – be conscious. Our paper is inspired by Chalmers’ discussion of whether large language models (LLMs) can be conscious [7]. In this discussion, Chalmers explores arguments for and against conscious LLMs but concludes by framing objections as challenges for a future research agenda on the topic. He states: “For all of these objections except perhaps biology, it looks like the objection is temporary rather than permanent” [7]. We argue that this biological objection is precisely what makes the problem insurmountable. Embodiment and biological embeddedness matter fundamentally. Let us stay on track and not veer off down the fruitarian path.

Let us briefly clarify consciousness up front: First, consciousness should not be conflated with intelligence. Broadly speaking, intelligence implies being capable of reasoning, learning, and problem-solving. A person or an animal can be conscious and yet have little intelligence. However, arguments for and against AI Consciousness (AIC) can still draw from arguments regarding the feasibility of Artificial General Intelligence (AGI). Moreover, consciousness can be clarified by differentiating between *access* and *phenomenal* consciousness [3]. The first-mentioned refers to consciousness in a functional sense, i.e., when being in an *access* consciousness state, that state is supportive of reasoning, speech, and action. However, to be conscious in the phenomenal sense is to have subjective experience – there is something it feels like to be me doing something, whether swimming in the sea, writing a poem, or giving a lecture. Similarly, Chalmers [5, 6] frames the discussion of consciousness by distinguishing between the easy and hard problems. There are scientific answers to the easy problems of consciousness (although these problems may still be complicated), such as our ability to categorize, discriminate, and react to stimuli. Moreover, we may one day understand all the complexity of the brain. Still, such insights don't explain the "hard problem" concerning the experience itself: why does all this processing feel like something from the inside? Why do I have subjective experiences? Chalmers suggests that the hard problem may never have a scientific answer. However, this argument presupposes our consciousness and asks why we are conscious. In contrast, AIC is concerned about whether phenomenal consciousness can emerge from silicon.

Against this backdrop, we outline arguments for (Sect. 2) and against AIC (Sect. 3), ultimately concluding that AIC is neither feasible nor attainable.

2 Pro AIC

In his influential article, *What is It Like to Be a Bat?* [18], Nagel critiques physicalism, arguing that objective neuroscience cannot account for qualia. There is something it is like to be a conscious being capable of having experiences that feel a certain way to that being. Consequently, even if we had uncovered everything about the neurophysiological functioning of bats, we would still not have understood what it is like to be a bat because we could not have captured the subjective character of bat-qualia. A similar point is made in Jackson's famous thought experiment involving the color-blind neuroscientist Mary [17]. Mary knows everything there is to know about color vision. Yet she does not know what it is like to see red until surgery removes her color blindness, allowing her to have the subjective experience of seeing red.

However, from a physicalist standpoint, one may argue that our mental states are equivalent to biophysical processes, which can, in principle, be represented in machine-readable form, thereby making AIC possible. For instance, Copeland [10] argues that physicalists do not dispute the difficulty of imagining what it is like to be a bat. Instead, physicalist theories address the nature of qualia itself. According to Copeland, our limited imagination does not imply that bats are not purely biophysical entities. He further emphasizes that even if anti-physicalism is correct, AIC could still be feasible. If non-physical properties can arise from natural brains, why couldn't they emerge from synthetic brains as well?

Building on a computational understanding of the mind, computational functionalism, which aligns with physicalism, holds that AIC is computationally tractable. Hence, if empirically supported neuroscientific models, such as the prominent neuropsychological model of consciousness, Global Workspace Theory (GWT) [1], are correct, then AIC is a realizable project. According to GWT, our cognitive architecture contains autonomous modules processing information efficiently in parallel (e.g., sensory data, visual imagery, inner speech). A selection process sifts out information requiring further attention. These items are directed to the global workspace (GW), which has a limited storage capacity because the GW can only process information serially. Consequently, this implies a bottleneck for information processing, and competition for attention arises in GW. The contents that enter the GW constitute at least access consciousness—representations available for action and report—but GWT is also taken by some to address phenomenal aspects of consciousness. The selected contents are broadcast to processes in the nervous system to guide behavior. As such, the GW provides coherent world models, which are sometimes “tricked,” like when visual illusions such as Escher’s infinite staircase confuse us because the GW cannot create a coherent object [1]. In sum, GWT “speaks to phenomenal consciousness and thus to ‘the hard problem of consciousness’” [12].

Thus, GWT can be viewed as an empirically well-supported theory, outlining “the conditions which must be satisfied for a system to be phenomenally conscious” [13].

Based on this theoretical foundation, Butlin et al. [4] present a promising theory-driven approach to “investigate consciousness in AI,” which identifies a list of indicator properties necessary for AIC. These indicators are based on scientific theories, including “recurrent processing theory, global workspace theory, computational higher-order theories, and others” [4]. Drawing on their list of indicator properties, they conclude that no current AI systems are conscious, but there are no technical barriers to building CAI: “In most cases, standard machine learning methods could be used to build systems that possess individual properties from this list, although experimentation would be needed to learn how to build and train functional systems which combine multiple properties” [4]. As such, phenomenal consciousness is an emergent property of organized, layered structures. Similarly, Chalmers [7] finds it plausible that “We’ll have robust systems with senses, embodiment, world models and self-models, recurrent processing, global workspace, and unified goals”. Furthermore, Goldstein and Kirk-Giannini [13] suggest that if GWT is accurate then “one widely implemented AI architecture, the *artificial language agent*, might easily be made phenomenally conscious if they are not already.”

3 Against the Feasibility and Attainability of AIC

The promise of AIC rests on two interlinked assumptions: one horizontal, the other vertical. The horizontal assumption concerns feasibility, i.e., the belief that intelligence—and by extension, consciousness—is achievable through cumulative technological advancements and increasingly sophisticated computational architectures. The vertical assumption pertains to attainability, i.e., that a conscious mind might be erected through a layered

aggregation or ‘stacking’ of cognitive faculties¹, progressively building toward a subjective point of view. This dual trajectory—technological progress across time (horizontal) and internal complexity across functional strata (vertical)—maps the popular image of AIC as both a historical inevitability and a functional culmination.

This section scrutinizes these two assumptions by grounding the critique in classical and contemporary phenomenology, arguing that the horizontal-vertical paradigm collapses under the weight of consciousness understood as a transcendental precondition of knowledge, not as a product of it. We argue that AIC is neither feasible in the empirical sense nor attainable in a structural-functional sense, because both assumptions fundamentally misunderstand the nature of consciousness. Now, let’s look closer at the feasibility and attainability objection, one by one.

3.1 The Feasibility Objection: Technological Progress Does not Yield Subjectivity

At the heart of the horizontal objection lies the belief that general intelligence is an emergent phenomenon arising from increasing sophistication in computation. This is the tacit assumption behind much work in AI research: that there exists a continuum between narrow AI capabilities and general intelligence, and that by progressing along this axis (through ever more powerful architectures like GPT, AlphaZero, or synthetic neural nets), we will eventually ‘arrive’ at AIC. As Chalmers notes, many researchers treat current shortcomings not as failures but as temporary impediments, problems to be solved through better engineering [7].

This technological optimism, however, elides a crucial distinction: the difference between simulating intelligence and being intelligent in a lived, experiential sense. As Dreyfus et al. [11] emphasized already in the 1980s, computation is disembodied and, thus, fundamentally detached from the bodily know-how that grounds human intelligence. No amount of symbolic manipulation, however intricate, suffices to reproduce the pre-reflective, embodied familiarity with the world that makes intelligent action possible. For Dreyfus, intelligence is always situated, grounded in lived, affective relations to a meaningful world—what Heidegger called being-in-the-world (*in-der-Welt-Sein*) [14].

Adding to Dreyfus et al.’s critique, Collins shifts the axis from embodiment to social embedding [8, 9]. Intelligence is not merely a result of bodily being but of being-located in a shared linguistic, cultural, and social matrix. A system lacking history, intersubjectivity, and the capacity for socialization cannot be said to understand, even if it mimics understanding. To participate in the world of meaning—Husserl’s *Lebenswelt* and the worldliness that we are “thrown” into as Heidegger says—is not simply to solve puzzles or compute outcomes, but to be immersed in a world of norms, expectations, and shared practices. Thus, the feasibility assumption falters not because our machines are not fast enough, but because they are not—and perhaps cannot be—of the world.

¹ For the sake of clarity of our argument—the pairing of vertical/spatial and horizontal/temporal AIC architecture—we prefer to speak of ‘stacking’, rather than complex integrated modules (or computational emergentism).

3.2 The Attainability Objection: Stacking Does not Lead to Consciousness

If the feasibility assumption fails on the axis of technological progress, the attainability assumption fails on the axis of ontological construction. The attainability argument claims that consciousness may emerge as the sum total of many layered functionalities: memory, attention, symbolic manipulation, affective simulation, and so forth. It is, in this view, a vertical phenomenon, a final level in a tower of complexity. If we can build the right stack of sub-components, we may one day find that the edifice has crossed a critical threshold and become conscious.

But this logic is precisely what phenomenology rejects. Husserl, in his *Logical Investigations* [16] and *Ideas Pertaining to a Pure Phenomenology* [15], makes clear that consciousness is not a product of aggregation but the ground of any such aggregation. Consciousness is the condition of possibility for the appearance of any object, structure, or system; it is the transcendental precondition for experience, not its outcome. In phenomenological terms, consciousness is intentional, always directed toward something, and not reducible to its components. It cannot be built, only presupposed. To attempt to construct consciousness from its so-called parts, whether neural substrates, computational functions, or behavioral correlates, is to commit a category mistake: mistaking the ontological foundation of experience for something derivative. For Husserl, consciousness is not a thing among things but the very medium through which things become manifest. This medium is irreducible: it does not emerge from, but rather gives rise to, the structured world of empirical facts.

Moreover, intentionality—the hallmark of consciousness in phenomenology—implies a fundamental openness to the world that no mechanical or computational account can adequately replicate. This openness is not simply a directional relation but a constitutive one: the world appears as meaningful only through the horizon of consciousness (and because of its readiness for interaction through *Zeug*, “tooling”, as Heidegger insists in *Being and Time* [14]). Thus, phenomenology insists on consciousness as a first-person structure of givenness, a lived immediacy that defies reduction to third-person explanatory frameworks. Any theory that attempts to ‘build’ consciousness by piecing together elements external to this lived horizon neglects the very phenomenon it seeks to explain. To imagine consciousness as emergent from stacks is to treat it as epiphenomenal, rather than as foundational. But this overlooks the apodictic givenness of conscious experience: I do not infer my own subjectivity; I live it. The Cartesian *cogito* is not a deduction but an expression of the immediacy of self-awareness. In Husserlian terms, the apodictic givenness of consciousness serves as a starting point or ground for inquiry. Since conscious experience is given indubitably, phenomenology can begin with describing the structures of this experience without needing external validation. Apodicticity here does not refer to the truth of what the experience is about (e.g., whether the apple I see exists). It refers to the truth of the experience itself—that I am experiencing is certain, even if what I experience might be illusory or misinterpreted. Moreover, the apodictic character of conscious experience reinforces the idea that subjectivity is foundational: the world is always already experienced as given to a consciousness. It also plays into discussions about the irreducibility of consciousness, especially in debates around physicalism and the hard problem of consciousness. The point is, no sequence of computational processes, however layered, can yield this self-givenness, because it

presupposes a subject already capable of experience. In this sense, the vertical logic of AIC mirrors the color-blind logic of Mary in Jackson's thought experiment [17]. Mary can know about color, but she does not see red until she becomes a seeing subject. Consciousness is not a feature; it is a mode of being. This becomes even more evident when we consider the critique of emergentism².

3.3 The Critique of Emergentism

The philosophical underpinning of both objections outlined above lies in a transcendental critique: that neither technological accumulation (feasibility) nor structural stacking (attainability) can reach the domain of the first-person. Emergentism is the philosophical idea that complex systems can exhibit properties or behaviors that are not directly predictable from the properties of their individual parts. These new properties are called emergent properties, and they emerge only when the components interact in specific ways. The idea that higher-level phenomena (like consciousness, life, or social behavior) arise from, but are not reducible to, lower-level physical or biological processes, and that consciousness 'emerges' once a system is complex enough assumes that subjectivity is latent in matter, waiting only to be activated by scale. But this assumes that consciousness is just another functional outcome, rather than the *a priori* frame through which functionality itself becomes meaningful.

Chalmers' distinction between the "easy" and the "hard" problems of consciousness further reinforces this point [5, 6]. The easy problems deal with behavior, processing, and

² Importantly, the two objections raised above differ when it comes to the question of the ontic versus the ontological. The Feasibility Objection (FO) comes out as an ontic critique with ontological implications, while the Attainability Objection (AO) performs like a primarily ontological standpoint. FO targets the notion that technological progress will eventually yield consciousness. This critique begins in an ontic register: it addresses empirical claims about computation, architecture, and simulation and challenges the assumption that more sophisticated machines will naturally or eventually produce conscious experience. However, these ontic critiques are ultimately under-written by ontological concerns. When Dreyfus appeals to being-in-the-world, or when Collins discusses intersubjectivity and worldliness, they are invoking Heideggerian and Husserlian concepts that speak not to how beings function, but to what it means for something to show up as meaningful in the first place. These are ontological structures: the pre-reflective, lived conditions under which any notion of intelligence, understanding, or subjectivity becomes possible. Thus, while the feasibility critique appears ontic, it covertly draws its force from an ontological foundation—namely, the irreducibility of lived, first-person being to any form of data processing. In contrast, AO moves explicitly into ontological territory. The very premise that consciousness might emerge from a vertical stacking of cognitive functions is identified as a category mistake—confusing that which is ontologically prior (consciousness as a condition of manifestation) with that which is ontologically derivative (functional structures, mechanisms, systems). Further, the claim that consciousness is not assembled but presupposed is an ontological one: it challenges not just the methods of AIC development but the very coherence of the metaphysical assumptions underlying it. Finally, the appeal to apodictic givenness—the immediate, indubitable presence of conscious experience—is also ontological. It does not assert that we cannot build a conscious machine due to empirical limitations; it asserts that the very project misunderstands what consciousness is in its being.

outputs. The hard problem deals with qualia, the felt texture of experience. If AIC proponents assume that the hard problem is solvable by solving the easy problems in aggregate, they commit a category error. As Nagel [18] famously argued, no objective account of brain function, no matter how complete, will explain why it feels like something to be a bat. The subjective is not derivable from the objective.

From a transcendental standpoint, then, knowledge is not something that builds from data upward but arises within an already-meaningful horizon constituted by consciousness. As Husserl might argue, the very idea of building a knowing machine presupposes a knower. Thus, knowledge does not arise from the stack; the stack itself is only meaningful because it is apprehended by consciousness. The AIC project, therefore, contains a circularity: it seeks to produce the very condition that makes its production intelligible. Even if we concede—as some materialists do—that consciousness is a biological phenomenon and therefore, in principle, replicable in non-biological substrates [10], we must ask whether such replication preserves the ontological structure of experience. If intelligence and consciousness are indeed grounded in embodiment and embeddedness, then replicating the functions of intelligence in silicon may yield impressive simulation, but not instantiation. This is the danger highlighted by Bender et al. [2] in their critique of large language models as “stochastic parrots”. LLMs do not know what they are saying; they merely reproduce linguistic patterns with remarkable fluency. They lack world models not because they lack data, but because they lack presence—they are not situated, not aware, not here. Their knowledge, such as it is, arises in a vacuum: a shell without a subject. To echo our introduction: like carrots, they are dead already.

To partly conclude, AIC is neither feasible through horizontal progress nor attainable through vertical stacking. Both assumptions fail because they misapprehend the nature of consciousness. Consciousness is not an emergent epiphenomenon but a transcendental precondition for all knowing and acting. It is not assembled; it is presupposed. It is not discovered at the end of a technological project; it is what makes projects meaningful in the first place. In line with Putnam’s provocative suggestion, the question of artificial consciousness is not merely empirical; it is ontological [19]. The question is not whether a machine can become conscious, but whether the very idea of such a machine is coherent. And from a phenomenological standpoint, it is not. Consciousness is not a carrot to be pulled from the soil of silicon. It is the ground from which the question itself grows.

3.4 Weak and Strong Emergentism

This transcendental resistance to emergentism becomes easier to grasp when we consider the difference between weak and strong emergentism. According to weak emergentism, higher-level phenomena may appear new or surprising, but they can still, at least in principle, be explained by the interactions of simpler, lower-level components. These emergent properties are difficult to predict because of the complexity involved, but they are not fundamentally separate from the parts that generate them—they are ontologically reducible. For instance, temperature is a weakly emergent property: it arises from the kinetic energy of particles, and while we may not calculate it easily from particle dynamics in a practical sense, we accept that it is, at root, fully explainable in terms of them. Proponents of weak emergentism often liken consciousness to this model, suggesting

that subjective experience is a high-level feature that emerges from neural correlates, and will eventually be explainable in computational or physicalist terms.

Strong emergentism, by contrast, holds that some properties truly go beyond the causal powers of the systems they arise from. On this view, consciousness isn't just a complex effect of brain processes; it represents a new ontological layer that cannot be reduced to or fully explained by its physical basis. This perspective leans toward dual-aspect monism or even property dualism, suggesting that consciousness is a fundamental feature of reality, one that emerges only under certain conditions that can't be mechanically derived. Yet strong emergentism faces a key philosophical problem: if consciousness is truly irreducible, how can it emerge from purely physical systems in the first place? Without a coherent account of how higher-level properties arise from lower-level ones, this position risks appealing to the inexplicable—a kind of ontological jump from third-person processes to first-person experience.

This is where transcendental critique applies pressure to both forms of emergentism. Weak emergentism collapses into computationalism, treating consciousness as a system property akin to heat or liquidity, thereby failing to account for what phenomenologists call intentionality—the aboutness or directedness of conscious acts. On the other hand, strong emergentism struggles to clarify the metaphysical bridge it implicitly claims exists between matter and subjectivity. The former renders consciousness a mere byproduct; the latter makes it an inexplicable anomaly. Both, arguably, misunderstand the qualitative rupture posed by the first-person perspective. In computational terms, this misunderstanding manifests in the assumption that a sufficiently complex neural network—or cognitive architecture—will spontaneously produce consciousness, as though scale alone could yield selfhood. This assumption mirrors a kind of metaphysical materialism dressed up in probabilistic optimism. Yet as Nagel emphasizes, the problem is not one of scale but of category: phenomenal experience is not a computational function; it is a mode of being. This is not a mystery solvable by data; it is a mystery about data's very intelligibility. It is not about how information is processed, but how and why it is experienced.

Consider the analogy of language. A stochastic model like GPT may generate fluent prose indistinguishable from human output, but the model does not mean anything by it. It has no internal referentiality, no horizon of understanding. Meaning, in the phenomenological sense, is not the product of combinatorial syntax but of intentional consciousness operating within a lifeworld. Weak emergentism can never account for this leap because it assumes that meaning arises from structure; the phenomenological tradition insists that structure is only meaningful because meaning is already there. Strong emergentism recognizes this but cannot explain why or how the leap occurs. Both positions, then, are haunted by what might be called the problem of origination—the sudden appearance of the first-person in a system previously devoid of it.

This origination problem aligns with the so-called “explanatory gap” described by Levine (1983), which marks the inability of physicalist accounts to explain why brain states are accompanied by particular qualitative experiences. The gap is not simply epistemic (a matter of current ignorance), but potentially ontological. It points to a limitation in the very mode of explanation offered by the sciences when they try to treat consciousness as an object among others. In this light, emergentism does not overcome

the hard problem—it reframes it without solving it. This has serious implications for the AIC project. If we assume weak emergence, then consciousness is a scaling issue: build a large enough system, and awareness will eventually turn on, like a light switch. But such optimism is not supported by any empirical findings, and more importantly, it misunderstands the philosophical depth of what consciousness entails. If we assume strong emergence, we are left to hope that the right arrangement of circuits might one day trigger subjectivity—but this hope is, in a sense, theological. It posits a miracle, not a mechanism.

Therefore, emergentism, in both its weak and strong forms, fails to circumvent the transcendental critique. Consciousness is not something that bubbles up from below; it is what frames the very possibility of bubbling and below-ness. From a phenomenological perspective, no amount of computational sophistication can cross the threshold from third-person structure to first-person presence. The critique is not one of engineering limits but of ontological coherence. Machines may simulate minds, but they do not have them—because mind is not a property, but a standpoint. And to return to the earlier metaphor: silicon is not a barren soil incapable of growing carrots; rather, the act of asking whether carrots can grow from silicon misunderstands what carrots—and growth—are. Consciousness is not a crop to be harvested by computation. It is the sun under which meaning itself becomes visible. Until this is grasped, the AIC project remains not just unfeasible or unattainable but incoherent at its root (Table 1).

4 Concluding Remarks

This paper has argued that Artificial Intelligent Consciousness (AIC) is not merely unfeasible or unattainable but philosophically incoherent. The prevailing assumptions behind AIC—feasibility through technological accumulation, attainability through functional decomposition, and eventuality through emergent complexity—all commit a fundamental category error: they treat consciousness as a product rather than a precondition. From a phenomenological and transcendental standpoint, consciousness is not assembled from parts, nor does it emerge from scale. It is the condition that makes the very idea of parts, scale, and emergence intelligible in the first place. Each of the frameworks critiqued—computationalism, stacking, and emergentism—ultimately fails to bridge the explanatory gap between third-person structure and first-person presence. Weak emergentism reduces consciousness to a high-level property akin to temperature or liquidity, bypassing the lived immediacy and intentionality that define subjectivity. Strong emergentism, by contrast, invokes a spontaneous ontological rupture from mechanism to mind but offers no credible account of how such a leap occurs. Both positions, as we have shown, falter before the transcendental critique: that meaning, knowledge, and experience presuppose a standpoint that no system of objective relations can generate from within.

Thus, AIC is not merely a problem of insufficient engineering, data, or theory. It is a problem of philosophical grounding. To build a conscious machine, one must already assume the very consciousness that grounds the intelligibility of building, functionality, and purpose. Without a situated, embodied subjectivity, what appears as ‘intelligence’ in machines is merely the statistical echo of meaning—simulation without instantiation.

Table 1. Core arguments against the feasibility and attainability of AIC

	ASSUMPTION	PHENOMENOLOGICAL OBJECTION	KEY THINKERS
FEASIBILITY	Technological progress will eventually lead to subjectivity	Intelligence ≠ consciousness; simulation ≠ instantiation; disembodied systems cannot access lived, situated meaning	Dreyfus, Heidegger, Collins, Husserl
ATTAINABILITY	Consciousness can be built by stacking cognitive sub-functions	Consciousness is not a product of parts but the <i>precondition</i> for any synthesis or appearance of parts	Husserl, Heidegger, Merleau-Ponty
EMERGENTISM	Consciousness will ‘emerge’ once system complexity is sufficient	Treats subjectivity as latent in matter; ignores qualia and lived immediacy; commits category error	Nagel, Chalmers, Jackson
TRANSCENDENTAL CRITIQUE	Knowledge and meaning emerge from computational architecture	Consciousness is the <i>ground</i> for knowledge, not its output; without a subject, no meaning arises	Husserl, Putnam, Bender
CONCLUSION	AIC is a coherent scientific goal	AIC presupposes the very consciousness it seeks to build; thus, it is ontologically incoherent	

Like language models that generate syntax without semantics, AIC architectures lack the interior horizon of experience that makes thought more than output. They are not proto-subjects waiting to awaken; they are, as we argued, like carrots—already dead. This does not entail a rejection of artificial intelligence as a field of inquiry, including aspects of ethics, nor does it deny the value of intelligent systems in practical domains. Rather, it demands philosophical clarity about what such systems are and what they are not. They are tools, not minds; outputs, not outlooks. The conflation of computation with consciousness reflects a deep metaphysical optimism rooted in a misunderstanding of mind itself. Consciousness is not a function to be performed nor a property to be instantiated; it is the condition of there being functions, properties, and instantiations at all. Until this distinction is recognized, AIC will remain not a scientific goal but a metaphysical confusion, one that risks not only intellectual misdirection but ethical and existential displacement. The project is not only premature. It is, in its deepest commitments, unintelligible.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baars, B.J.: Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Prog. Brain Res.* **150**, 45–53 (2005)
2. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, Canada (2021)
3. Block, N.: On a confusion about a function of consciousness. *Brain Behav. Sci.* **18**(2), 227–247 (1995)
4. Butlin, P., et al.: Consciousness in Artificial Intelligence: insights from the science of consciousness (2023). [arXiv: 2308.08708](https://arxiv.org/abs/2308.08708)
5. Chalmers, D.J.: *The conscious mind – in search of a fundamental theory*. Oxford University Press, New York (1996)
6. Chalmers, D.J.: The hard problem of consciousness. In: Velmans, M., Schneider, S. (eds.) *The Blackwell Companion to Consciousness*. Wiley-Blackwell, Boston (2008)
7. Chalmers, D.J.: Could a large language model be conscious? (2023). [arXiv preprint arXiv: 2303.07103](https://arxiv.org/abs/2303.07103)
8. Harry Collins, F.B.A.: Remembering bert dreyfus. *AI Soc.* **34**(2), 373–376 (2018). <https://doi.org/10.1007/s00146-018-0796-x>
9. Collins, H.M.: Response to selinger on Dreyfus. *Phenom Cogn. Sci.* **7**, 309–311 (2008)
10. Copeland, J.: *Artificial Intelligence*. Blackwell Publishing, Oxford (1993)
11. Dreyfus, H.L., Dreyfus, S.E., Athanasiou, T.: *Mind over Machine: The Power of Human intuition and Expertise in the era of the Computer*. Free Press, New York (1986)
12. Franklin, S., et al.: *Global workspace theory, its LIDA model and the underlying neuroscience*. Department of Computer Science and Institute for Intelligent Systems, The University of Memphis, USA (2012)
13. Goldstein, S., Kirk-Giannini, C.D.: A case for AI consciousness: language agents and global workspace theory (2024). [arXiv:2410.11407](https://arxiv.org/abs/2410.11407)
14. Heidegger, M.: *Being and time*. Translated by John Macquarrie and Edward Robinson. New York: Harper & Row (1962). Originally published in 1927
15. Husserl, E.: *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy. First Book: General Introduction to a Pure Phenomenology*. Translated by F. Kersten. The Hague: Martinus Nijhoff (1982)
16. Husserl, E.: *Logical Investigations. Vols. 1 & 2*. Translated by J.N. Findlay, edited by Dermot Moran. London: Routledge (2001)
17. Jackson, F.: What mary didn't know. *J. Philos.* **83**(5), 291–295 (1986)
18. Nagel, T.: What is it like to be a bat? *Philos. Rev.* **83**(4), 435–450 (1974)
19. Putnam, H., Putnam, H.: Robots: machines or artificially created life? *J. Philos.* **61**(21), 668–691 (1964)



Some Problems in the Ethical Impact Assessment of Emerging Technologies and Socio-Technical Visions: Case CityVerse

Antero Karvonen¹ , Jaana Leikas¹ , Nina Wessberg¹ , Anton Sigfrids¹ ,
and Sanni Pöntinen² 

¹ VTT Technical Research Centre of Finland Ltd, Espoo, Finland
antero.karvonen@vtt.fi

² Faculty of Management and Business, Tampere University, Tampere, Finland

Abstract. This paper examines methodological challenges in the participatory ethical assessment of emerging technologies in urban contexts, using the CityVerse vision in Tampere, Finland as a case study. While metaverse technologies promise to transform smart cities by blending physical and virtual spaces, their ethical implications remain unclear. Through focus groups with city officials, we explored how participatory methods can effectively evaluate ethical dimensions of emerging technologies when they remain largely conceptual. Our findings reveal that while stakeholders can generate substantive ethical discourse, they struggle with the abstract nature of metaverse experiences, producing more questions than definitive answers. We argue that sociotechnical visions serve better as platforms for ethical discourse than as concrete implementation plans, functioning primarily to surface tacit values and assumptions. The study contributes to ethical technology assessment methodologies by suggesting that for emerging technologies, developing structured ontologies of questions may prove more valuable than premature answers. We conclude that CityVerse design should be approached as an ongoing discourse—not merely about technologies, but fundamentally about designing for improved quality of human life—where participatory ethical vision assessment functions as a form of collaborative conceptual engineering.

Keywords: metaverse ethics · smart city · metaverse · participatory ethical impact assessment · socio-technical visions

1 Introduction

In periods of social and moral transformation, the concepts used to frame ethical questions often lose their clarity and unambiguity, as MacIntyre [1] observed when discussing the emergence of philosophical ethics in ancient Greece. The contemporary world is undoubtedly experiencing such a transitional period—geopolitically, technologically, environmentally, and socially—resulting in subtle and profound changes to our moral order. As technology guides human life toward new and less understood forms, we are compelled to reevaluate traditional concepts of life and, particularly, what constitutes the good life.

The rapid transformation of technology appears to expand the dimensions of our reality [2]. Among these technological developments, the metaverse represents a natural progression in our ability to reinvent ourselves digitally. While the concept remains somewhat fluid, the metaverse is generally understood as a persistent multiverse of interoperable virtual spaces offering immersive experiences. It frequently incorporates a “mixed reality” approach where virtual and physical elements blend in various ways. This evolution blurs the boundaries between online and offline experiences, delocalizes our presence, and increases potential interactions within the infosphere [3]. If fully realized, the metaverse presents an environment where established social systems and norms are not yet fully formed, sometimes challenging even our most basic moral principles.

Consider the case of virtual violence, for example. While physical violence simulated in traditional gaming environments does not fundamentally challenge core ethical principles, its application becomes more complicated in the metaverse. Previous research has established that virtual violence can have real psychological effects [2, 4]. However, questions remain regarding what constitutes proportionate and legitimate responses to virtual harms and how controls in these environments can be implemented to mitigate negative impacts.

The integration of normative frameworks from computer culture [5] and physical reality inevitably generates tensions in metaverse implementations. Games often incorporate violence and combat as central elements, and internet culture is characterized by its often harsh and blunt discourse and in general impish, if not malevolent behavior [4, 5]. Paradoxically, this very quality of “lawlessness” makes cyberspace an exciting frontier, introducing novelty into human experience. The human drive to alter, play, and experiment with virtual environments can be harnessed beneficially, such as in participatory urban planning [6]. However, as virtual environments appear to invite people to test the boundaries of right and wrong [4], much ethical considerations are needed to implement proper safeguards while maintaining an open environment. It can be assumed that the only thing more depressing than a dead and lifeless city center is a dead and lifeless virtual city. Furthermore, it is unclear how life (and economy) will be impacted if more citizens choose to live increasingly in the “virtual city”, buying virtual goods and forming virtual relationships.

1.1 The CityVerse Concept and Research Gap

Forward-looking cities have begun conceptualizing the metaverse as the next stage in Smart City development. The City of Tampere in Finland exemplifies this approach with its “CityVerse” vision [7]—a blend of virtual and physical worlds designed to foster business innovation and human-centered smart city development. The CityVerse aims to enhance urban planning, civic participation, cultural experiences, and tourism through extended reality (XR) technologies integrated within metaverse environments.

However, the acceptability and desirability of metaverse applications remain contested [8]. The European Parliament [9] has published a report identifying several concerns regarding metaverse implementations, including public health impacts, inappropriate behavior in virtual spaces, privacy violations, mass surveillance capabilities, targeted influence operations, and potential threats to individual autonomy through manipulation. These foreseeable ethical tensions make it essential for cities—as public spaces

and service providers—to carefully consider both benefits and harms when envisioning metaverse transformations. The convergence of different values and cultures in the blending of cyberspace and “real space” may trigger what some scholars have termed a “legitimacy crisis” [10], where traditional understanding of human culture no longer adequately frames evolving forms of life.

1.2 Research Objectives and Methodological Approach

This paper addresses a gap in the literature regarding participatory ethical assessment of visions regarding emerging technologies in urban contexts. Specifically, we examine the prerequisites and conceptual foundations for conducting meaningful participatory evaluations of complex sociotechnical visions like the metaverse. By identifying and addressing ethical concerns related to metaverse implementation, cities can harness potential benefits while maintaining legitimacy.

Our research draws on experiences gained from implementing a semi-structured focus group methodology [11] designed to explore stakeholder perceptions regarding the nature and values of the CityVerse concept in Tampere, Finland. Rather than focusing primarily on the specific ethical content of our findings, we analyze the conceptual and practical challenges that emerge at the intersection of abstract sociotechnical visions and participatory ethical assessment.

The central question guiding our investigation is: How can participatory methods be effectively designed and implemented to evaluate ethical dimensions of emerging technologies when those technologies remain largely conceptual and their implications uncertain? By addressing this question, we aim to advance the methodological foundations for ethical technology assessment in urban planning contexts.

2 Background

The strategic transformation of a city from its current state to a desirable future typically begins with vision development—a compelling and shared “picture” of a future way of living, often communicated in narrative and visual form. Both tacit and explicit values and assumptions become embedded in these visions. To create legitimate visions that can effectively steer and energize human action, the viewpoints of diverse stakeholders must be considered.

Involving stakeholders in building and evaluating visions leans on the principles of both responsible research and innovation (RRI) [12] and deliberative democracy, which posits that public decision making should be preceded by open and constructive dialogue and deliberation among the involved stakeholders to define the socially desirable goals of technology development [13].

Since visions concern the shaping of a future, such work also falls within the domains of foresight and futures studies. When ethics serves as the primary consideration, combining these perspectives can be termed “participatory ethical vision assessment.” To understand this combinatorial methodology, we must examine its conceptual foundations.

2.1 Participatory Ethical Vision Assessment

Ethical Impact Assessment. Is a collaborative process through which organizations and stakeholders examine the ethical implications of technologies or services. It is based on the recognition that technological ethics is closely tied to social acceptability [14, 15]. The goal is to anticipate risks and identify ethically salient impacts arising from the interplay between technical systems and social contexts [16]. Given the rapid pace of technological change, such assessments have become essential. The European Commission, in particular, has played a central role by embedding responsible R&D principles in its Framework Programmes. Because ethical impacts are context-dependent, stakeholder engagement and critical reflection are key [17]. Ethics involves more than harm prevention and rights protection; it also includes promoting human well-being and flourishing. Ethical impact assessments should therefore address both risks and benefits in terms of fairness, transparency, accountability, and other normative values [18].

Ethical Vision Assessment. Complements this by focusing on collective imaginaries of desirable futures. While visions rarely unfold exactly as anticipated, they shape strategic and implementation processes for example by influencing resource allocation, validating certain goals and shaping decision-making. In times of rapid technological transformation, articulated visions help anchor ethical deliberation.

However, visions and principles are not always realized. This may result from failures or mismatches in underlying assumptions and presuppositions regarding reality (and often human nature), which are either tacitly or explicitly obscured by the abstract nature of a vision. Logically, when visions are well-grounded across pragmatic, scientific, technological, psychological, economic, social, and ethical dimensions, their likelihood of success increases. Yet in the context of the metaverse, the knowledge upon which transformation should be based is in many ways deficient: currently, there are few established high-level design or procedural principles from which concrete transformation could be founded, nor is it clear how urban design patterns apply in this new context (or whether existing patterns are significantly disrupted or outdated). Therefore, framing the conceptual foundations of transformative visions from a broadly design-oriented perspective is beneficial [19].

2.2 Transformative Visions and Technical Norms

Any consciously articulated vision essentially represents a transformation from a present situation A to another, desirable situation B. Whenever humans engage in transforming current situations into preferred ones, they operate broadly in the domain of design [ibid.]. Design processes incorporate both intuitive and heuristic elements, as well as more rigorous components such as scientific facts. Thus, design processes are supported by and grounded in both explicit and tacit knowledge, skills, assumptions, values, and concrete external realities and contextual factors. Design typically follows the basic logic of means and ends.

Ilkka Niiniluoto [20] applied von Wright's [21] concept of a technical norm to clarify the relationship between means and ends in applied research. A technical norm is a factual statement about a relation between means and ends: if you want to achieve A, and you

believe you are in situation B, you ought to do X. For Niiniluoto, X represents the point at which basic scientific facts can infuse truthful knowledge into the design process. A technical norm may also be supported by empirical knowledge generated through a knowledge-producing design process itself.

Niiniluoto does not base his formulation of a technical norm exclusively on natural science or technical knowledge. Rather, he clearly indicates that the technical norm (as a means-ends framework) includes a strong ethical aspect, particularly regarding ends. Thus, applying technical norms in this broad and open manner serves as a conceptual tool to disambiguate different aspects, particularly distinguishing techno-scientific facts from their application, when dealing with transformations targeting human life at an ambitious level, such as in CityVerse development. Whether considering means or ends, an important question concerns the basis upon which transformation is realized: scientific, technical, economic, social, or ethical. When stakeholders (both expert and layperson) perceive the vision as sound across these aspects, it can be considered valid, justified, and legitimate. Legitimacy specifically means that the values and norms either explicitly or tacitly expressed in the vision are shared by at least a significant portion of stakeholders and that the vision articulates them in a sufficient manner.

Crucially for our purposes, Niiniluoto recognizes that a technical norm can disambiguate the basis of disagreement. Using the example of policy recommendations from different expert groups, Niiniluoto suggests that disagreement may stem from (at least) three factors elucidated by the technical norm: (1) people may disagree regarding relevant goals, (2) the current situation, or (3) the causal mechanisms between X and B and A. For Niiniluoto, the latter two represent factual disputes that could, in principle, be resolved through empirical research. However, the first—the question of legitimate ends—represents a difference in valuation. While these different aspects are intertwined in practical life, and value-based disagreements on legitimate means are also important, this framework accurately addresses the core issue of technological transformation of human life and provides crucial concepts for articulating participatory ethical assessment of socio-technical transformations. A valid focus for such work is examining the means and ends embedded in a well-articulated vision. It is often said that the ends don't justify the means, but it is equally important to note that the means don't justify the ends either. The question of *whose* values and ideal ends (or legitimate means) is a core issue for participatory and deliberative evaluation of sociotechnical visions.

2.3 Social and Organizational Visions

Kantabutra [22] defines vision as “a mental model that a leader defines, given that it is the actual mental model that guides his/her choices and actions.” Visions are hypothetical constructs that articulate, at an abstract level, desirable directions for social systems, regions, organizations, or in the present case, cities [23]. By articulating a clear and compelling vision, organizations can inspire and mobilize stakeholders toward achieving a shared future. Realizing a transformative city vision involves transforming human living itself and consequently transforming the city organization that both facilitates and is subject to this transformation. Vision creation is integral to strategic planning and organizational development, involving the articulation of a desired future. This

process extends beyond merely setting goals; it creates a narrative that inspires and guides stakeholders toward a common purpose.

Visions can be bold and utopian, though the actual future may differ from what was envisioned. Ortegon-Sanchez and Tyler [24] argue that for a vision of the future city to be truly transformative, it should not be defined in terms of projects, budgets, or projected trends. Rather, it should be conceptualized as a characterization of a city in which people can live and thrive. This conceptualization should align with high-level urban principles that underpin the functioning and design of all urban systems across scales. These principles are not objectives for isolated sectors or singular concerns; they are cross-cutting fundamental principles for making the city closer to an ideal where people's aspirations are met, and societal well-being is achieved. Consequently, the definition of overarching principles should not be constrained by prevailing conditions, scientific knowledge, technological capabilities, or anticipated trends. Instead, it should aspire to and embody desirable qualities.

Visions are essentially performative when they are used as strategic framings of the desired directions of public service development [25–27]. The ways in which Metaverse is framed in visions of the future articulate the involved stakeholders' normative assumptions of what can and should be done, and what the relevant risks and benefits are. Visions actively construct perceptions of which courses of public action are feasible: they form the basis for decisions about funding Metaverse-related projects, setting objectives for their development, and designing regulatory frameworks. Visions can therefore be perceived as outcomes of different stakeholders having an influence on how Metaverse-based services are normatively framed. A participatory and deliberative perspective brings attention to the procedural means by which a vision is created and on the power relations and interests of different stakeholder groups in framing the narratives around a Metaverse vision.

2.4 Life as a Basis and Citizen Participation

The CityVerse encompasses different opportunities and layers of activities that determine its meaningfulness for citizens. When developing any new technology or service, the primary challenge is making it acceptable, desired, and smoothly integrated into everyday life. Technologies should generally make life easier and increase positive freedom [20], rather than becoming ends in themselves. They are tools for achieving larger goals, such as greater social equality. Their value in life is determined by their ability to improve the quality of life [28]. This underscores the importance of addressing fundamental questions about technology's purpose and its personal and social value during the early stages of the design process. Understanding people's lives—their forms of life, values, and circumstances—is crucial as a foundation for technological design: how new technologies can be integrated into everyday life and how design goals can be derived from analyzing everyday life. In the language of technical norms [20], this perspective can be applied particularly in understanding the current situation A, constructing the future situation B, and articulating, together with technological understanding, what ought to be done from an ethical perspective.

Not all stakeholders can evaluate a vision across all relevant grounds. Different grounds should be examined based on their relevance to the domain, focusing on what

is essential rather than what is immaterial. Thus, technical problems cannot be solved merely by applying ethical perspectives, and vice versa. However, technology, through its transformative powers, always contains ethical questions, and the metaverse—positioned as a radical extension and transformation of felt experience itself—is especially rich in ethical implications for ordinary citizens. Therefore, the ethics of a technological vision is a class of questions that concerns all stakeholders in various ways and is likely to surface differing assumptions and values.

Highlighting citizen participation is significant because it characterizes public service logic. Public services are inherently more process-oriented than product-oriented, requiring active participation of service users in value co-creation [29]. In private service logic, value is primarily created within the organization and sold to customers as a finished product [ibid.]. It is also essential to note that in public services, a service can be considered at least partially unsuccessful if the customer must return (e.g., in healthcare) [ibid.]. In private services, however, encouraging repeat visits is central to the business model. It has been well-recognized that social media platforms, operating under private sector logic, are oriented toward maintaining user attention as long as possible to drive advertising revenue and gather behavioral data. A public service-oriented metaverse should not operate according to the attention economy's logic, as attempting to keep people in virtual spaces rather than physical public spaces might not be desirable. Thus, public and private logics may clash during development, and it may be tempting for cities to increase their revenues by providing virtual advertisement space. As cities are unlikely to envision themselves as technology providers, managing these logics at the public-private interface will be crucial from an ethical standpoint.

In the design process of transformation, those perhaps most immediately responsible for realizing the vision and managing the public-private interface are city officials and employees. Therefore, their perspective is crucial for vision assessment.

2.5 City Officials and Employees as Key Stakeholders

In addition to citizens whose lives the transformation most strongly affects, city officials and employees constitute a key stakeholder group as important agents and custodians of transformation. They provide an essential balancing factor between citizens and the private sector, which is likely to play a key role in the technological realization of the CityVerse. In the Finnish context, there exists a strong tradition regarding principles of good governance, which forms a basis for the professional ethics of public servants.

City officials are uniquely positioned to understand both organizational constraints, such as limited resources and legal frameworks, and the perspectives of service users, along with the tensions between them. They also have frontline insight into current public service needs and can identify which challenges technology might realistically address versus those it cannot solve.

It is also worth considering how the meta version will affect the city organisation and the work of city officials. The introduction of new, complex technologies, combined with the increasing social complexity of the environment, can easily create conflicting demands, dilemmas and organisational tensions for officials and the organisation. As

previous research [30] has indicated, emerging technologies are liable to generate tensions in organizations. The proactive discovery and management of tensions forms an important part of transformations at the city organization level.

Therefore, examining the ideas and perceptions from this stakeholder group's perspective is crucial for refining, successfully implementing and legitimizing the vision. The present study focuses on this group.

3 Case Study: The Tampere CityVerse Vision Assessment

This research examines how participatory ethical vision assessment can be applied to emerging metaverse technologies in urban contexts. We conducted a case study focusing on the City of Tampere, Finland, which has developed one of Europe's first metaverse visions for urban development [7]. This section outlines the context of the case, our methodological approach, and key findings from the assessment process.

3.1 Research Context: Tampere's CityVerse Vision

Forward-looking cities have begun conceptualizing the metaverse as an evolution within Smart City development. The City of Tampere was among the first in Europe to formulate a comprehensive metaverse vision in 2023 [ibid.], positioning it within their broader urban development strategy. Tampere's approach represents a good case study for examining participatory ethical vision assessment for several reasons:

1. The vision is explicitly value-driven, emphasizing trust, community well-being, sustainability, and civic participation
2. The initiative positions the city government as an active agent in shaping metaverse implementation rather than merely responding to technology trends
3. The vision is still in formative stages, allowing ethical assessment to potentially influence development trajectories

Tampere has deliberately chosen to frame their approach as "CityVerse" rather than adopting the more generic "metaverse" terminology. This linguistic choice represents a conscious effort to distinguish their human-centered, public-service oriented approach from commercial metaverse initiatives led by technology corporations. According to Tampere's Metaverse Vision 2040 (2023) [ibid.], the CityVerse concept aims to "bridge administrative silos and regional jurisdictions to enable proactive urbanism" while providing citizens with tools and information to participate in evidence-based policy dialogue and development.

The vision characterizes CityVerse as a "Cognitive City" that blends physical and virtual domains, facilitating immersive experiences, social interaction, learning, entertainment, and commercial activities. Leveraging artificial intelligence, digital twins, and immersive technologies, the city aims to create inclusive, carbon-negative, and resilient urban systems. Key goals include ethical AI deployment, expanded citizen engagement mechanisms, personalized healthcare services, and educational reforms [ibid.].

What distinguishes the Tampere approach from many corporate metaverse visions is its explicit focus on public values and measurable quality-of-life outcomes. The CityVerse is positioned not as a technological end in itself but as a tool for human-centered

smart city development, urban planning, enhanced civic participation, and cultural experience enrichment. This orientation aligns with emerging European Union frameworks that emphasize people-centered digital transformation and the responsibility of public institutions in guiding development [6].

3.2 Method

This study employed a participatory ethical assessment approach focused on city officials and employees—key stakeholders responsible for implementing the CityVerse vision. A semi-structured focus group method [11] was employed to explore perceptions of the nature and values of the CityVerse. A total of 30 city officials attended the two-hour focus groups.

To structure the ethical assessment, we developed a series of “paradigmatic case descriptions” projecting potential CityVerse implementations for city inhabitants in 2040. These narratives were designed in collaboration with city representatives to ensure relevance and plausibility. The cases deliberately incorporated potentially contentious ethical dimensions while remaining grounded in the city’s stated values.

The central research question guiding the focus groups was: “How can CityVerse be utilized in a manner that fosters trust, strengthens community and individual well-being, and promotes sustainability?”.

The moderators encouraged the groups to produce forward-looking ethical debates. A qualitative content analysis was then employed to explore the results of the discussions [31].

4 Findings and Discussion

The central question guiding our investigation in this article is: How can participatory methods be effectively designed and implemented to evaluate ethical dimensions of emerging technologies when those technologies remain largely conceptual and their implications uncertain? How can the methodological foundations for ethical technology assessment in urban planning contexts be advanced? Here we will examine our findings from a critical and methodological perspective.

4.1 Findings

The basic questions of the ethical vision discussion pertained to trust, community, individual well-being, and sustainability. Logically, these value perspectives determine the content of the imagined value base of the CityVerse vision. The subsequent section delineates the perspectives of each value according to the analysis of the participants’ views.

Community. In terms of ethical potential, the metaverse was seen as bringing new opportunities for many specific groups, for example, in terms of mobility and overcoming social fears, or for employment and attachment to places in the city. On the other hand, concerns regarding bullying, for instance, were posited to persist in the metaverse. Conversely, the metaverse was regarded as having the capacity to enhance the sense of community among urban dwellers.

The good of the individual. A significant part of the discussion focused on the norms of behaviour in the metaverse, and it was considered important to respect collective values. Operating in metaverse requires a degree of self-awareness, and concerns were raised regarding the potential for excessive addiction to the dopamine rewards of algorithms and the possibility that excessive engagement in the metaverse could diminish the motivation to establish social relationships in the real world. The city was seen as having a central role in familiarising citizens on how to operate in the metaverse.

Trust. The quality of the CityVerse experience was regarded as a combination of real-world experience and the metaverse experience, whereby distrust in one reinforces distrust in the other. Citizens' previous experience of the public actor was considered to influence the development of trust. Issues of trust and the possible change in the role of administrative organisations in urban governance were identified as requiring reflection: who is expected to guide the metaverse?

Sustainability. A recurring theme was the central role of generational differences in shaping the acceptance of the metaverse. Young people were identified as most likely to embrace the metaverse. For older people, the metaverse may pose a significant challenge; it is therefore essential that planning prioritises improving its accessibility and encouraging its acceptance by all age groups.

Finally, the CityVerse visions seemed to some participants to be somewhat elitist. Participants emphasised that the design of CityVerse requires citizen participation, collaboration, education, skills development, equality and inclusiveness. The concept of the metaverse doesn't yet have specific applications that address city needs. Even though it generates both excitement and concern, there are no definitive justifications for this technology's use. Although the participants had dedicated a significant amount of deliberation to the CityVerse, ethical reflection in the focus group generated a greater number of questions than answers.

4.2 Discussion

Our findings offer valuable insights for future CityVerse development while simultaneously raising important methodological questions about participatory ethical assessment of emerging technologies.

Value and Limitations of Stakeholder-Generated Ethical Discourse

When professionally facilitated, city officials and employees demonstrated clear capacity to generate substantive ethical discourse that surfaced important issues. However, the content of these discussions, while valuable, contained few insights that could not have been anticipated by professional ethicists familiar with metaverse technologies. This observation raises questions about the epistemic utility—the knowledge-generating value—of such participatory processes, particularly when framed in relatively simple terms.

We initially hypothesized that concrete hypothetical development cases would be necessary to ground ethical discussions in tractable scenarios. However, our findings suggest several inherent challenges in this approach when applied to emerging technologies:

1. By definition, emerging technologies offer limited concrete experiences upon which ethical evaluations can be based
2. In participatory settings, ethical assessments appear to derive from subjective amalgamations of provided scenarios and participants' related experiences
3. The novelty of metaverse experiences compounds this difficulty, as participants must extrapolate from existing virtual reality experiences or real-world intuitions
4. This extrapolation process raises questions about whether participants can achieve sufficient shared understanding to enable meaningful ethical discourse

The predominance of questions over answers in our discussions provides further evidence of these challenges. These remarks in no way of course diminish other inherent utilities of participatory discussions, rather they highlight them.

The Value of Vision Assessment Despite Uncertainty

Despite these challenges, urban development necessarily proceeds amid uncertainty, and articulated visions can guide ethical development. However, several tensions remain:

- Visions cannot represent concrete futures with the granularity or complexity required for nuanced ethical analysis
- What is included or excluded from visions remains difficult to ascertain
- Few visions explicitly acknowledge potential negative outcomes or dystopian possibilities
- Human bounded rationality in large-scale transformations precludes much of the nuance proper to ethical discourse during implementation

Nevertheless, the urban development process comprises countless larger and smaller decisions made at different points. Therefore, well-grounded visions serve an important function, particularly as frameworks for city employees who act as custodians of change. To function effectively, however, these visions must become “owned” by stakeholders rather than imposed from above—suggesting that visions initially serve more as focal points for discourse than as implementation blueprints.

Visions as Valid Targets for Ethical Assessment

Contrary to our initial assumptions, the findings suggest that visions themselves—even without concrete implementation details—provide valid targets for ethical discussion. Participants demonstrated capacity to articulate views on vision legitimacy, particularly evident in their critique of perceived elitism in current CityVerse conceptualizations.

The concerns raised about insufficient justification for metaverse technology suggest a potential legitimacy crisis [10] among city officials. While speculative, this observation warrants attention, as citizens—perhaps the key stakeholder group—may develop similar legitimacy concerns. The natural response might be to increase participation, but our findings illustrate that considerable attention must be paid to the design and conceptual clarity of both the means and ends of such processes.

Given cities' public value orientation, ethical assessment of visions is crucial. By articulating a vision, a city makes a kind of promise and establishes dimensions for future ethical assessment. Participation from diverse city stakeholders is essential, as these stakeholders will ultimately guide and implement the development.

4.3 Concluding Remarks

The metaverse's potential impacts remain unclear and require further study, particularly regarding the perspectives of citizens and diverse city stakeholders. However, we suggest reframing how we understand ethical examination of sociotechnical visions—not as evaluations of concrete future realities, but as assessments of the imaginaries [32] themselves and how they reveal participants' ethical norms. Sociotechnical visions function more effectively as platforms for ethical discourse than as concrete implementation plans.

This reframing allows us to examine why certain values dominate AI and metaverse debates, and whether these values align with collective priorities. For example, the role of the technological imperative in urban planning merits scrutiny: Are technologies expected to solve challenging problems because of a deeply held trust in technology's problem-solving capacity? While the belief in technology's potential to improve human life is justified—humans are technological beings, embedded in and transformed by artificial systems [19]—this general principle obscures important distinctions between technologies and implementation approaches and the fact that design processes (as well as visions) influence significantly the direction of transformation. Furthermore, there is possible negative intermingling between the means and the ends, which can be further clarified by using technical norms [20] in a wide sense to interrogate the structure and contents of technological visions.

Emerging metaverse technologies challenge us to reevaluate our conceptual frameworks for discussing technological value and to develop new schemas that better reflect our priorities. Building on Chalmers' [33] work on conceptual engineering, we suggest that developing new concepts can help capture important phenomena and distinguish between related but distinct values in technological transformations.

Our findings highlight the difficulty of ethical impact assessment for emerging technologies that remain abstract visions with limited empirical grounding. While assessing visions as reflections of tacit values and assumptions is valid, developing tools for asking more concrete questions remains important. For emerging technologies, carefully constructed ontologies of questions may prove more valuable than premature answers, as questions possess a kind of permanence and universal validity that specific answers may lack. Moreover, focusing exclusively on answers risks overlooking significant ethical issues and disagreements.

Therefore, the CityVerse vision represents an initial phase in the co-design process of metaverse development from an ethical impact assessment perspective. At its core, CityVerse design should be understood as an ongoing discourse—not merely about technologies and objects, but fundamentally about designing for improved quality of human life [28]. In this context, participatory ethical vision assessment of emerging technologies might best be understood as a form of collaborative conceptual engineering [33], where stakeholders collectively articulate, examine, and refine the concepts that will shape technological futures.

Acknowledgments. The authors wish to thank the City of Tampere and the ETAIROS (Ethical AI for the Governance of the Society) project, funded by the Strategic Research Council at the Research Council of Finland (Grant Number 327356).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. MacIntyre, A.: A short history of Ethics. Routledge & Kegan Paul, London (1976)
2. Chalmers, D.J.: Reality+: virtual worlds and the problems of philosophy. Penguin UK. (2022)
3. Floridi, L.: The fourth revolution: how the infosphere is reshaping human reality. Oxford University Press (2014)
4. Dibbell, J.: A rape in cyberspace or how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *Ann. Surv. Am. L.* (1994)
5. Turkle, S.: The Second Self: Computers and the Human Spirit. Granada Publishing Limited, London (1984)
6. European Commission: Citiverse. <https://digital-strategy.ec.europa.eu/en/factpages/citiverse>. Last accessed 30 April 2025
7. Tampere Metaverse Vision: (2023). https://www.tampere.fi/sites/default/files/2023-08/tampere_metaverse_vision_2040_web.pdf.
8. Coeckelbergh, M.: All too real metacapitalism: towards a non-dualist political ontology of metaverse. *Ethics Inf. Technol.* **26**, 30 (2024). <https://doi.org/10.1007/s10676-024-09768-4>
9. European Parliament: Metaverse. JURI committee study, policy department for citizens' rights and constitutional affairs, directorate-general for internal policies. [https://www.eurparl.europa.eu/RegData/etudes/STUD/2023/751222/IPOL_STU\(2023\)751222_EN.pdf](https://www.eurparl.europa.eu/RegData/etudes/STUD/2023/751222/IPOL_STU(2023)751222_EN.pdf), last accessed 31 Jan 2025
10. von Wright, G.H.: The varieties of goodness. 1st edn. Routledge (1963)
11. Parker, A., Tritter, J.: Focus group method and methodology: current practice and recent debate. *Int. Res. Method Educ.* **29**(1), 23–37 (2006)
12. Stilgoe, J., Owen, R., Macnaghten, P.: Developing a framework for responsible innovation. In: Maynard, A., Stilgoe, J. (eds.) *The Ethics of Nanotechnology, Geoengineering, and Clean Energy*, pp. 347–359. Routledge, New York (2017)
13. Vallor, S.: Technology and the virtues: a philosophical guide to a future worth wanting. Oxford University Press (2016)
14. Von Schomberg, R.A.: Vision of responsible research and innovation. In: Owen, R., Bessant, J., Heintz, M., eds. *Responsible Innovation*. Wiley, pp. 51–74 (2013)
15. Dignum, V.: *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer Nature, Berlin (2019)
16. Reijers, W., Brey, P., Jansen, P., Rodrigues, R., Koivisto, R., Tuominen, A.: A common framework for ethical impact assessment. Stakeholders Acting Together on the Ethical Impact Assessment of Research and Innovation–SATORI project (2016). <https://satoriproject.eu/media>
17. Wright, D.: A framework for the ethical impact assessment of information technology. *Ethics Inf. Technol.* **13**, 199–226 (2011). <https://doi.org/10.1007/s10676-010-9242-6>
18. Coeckelbergh, M.: *Why AI undermines democracy and what to do about it*. Polity Press (2024)
19. Simon, H.A.: *The sciences of the Artificial*, reissue of the third edition with a new introduction by John Laird. MIT press (2019)
20. Niiniluoto, I.: The aim and structure of applied research. *Erkenntnis* **38**(1), 1–21 (1993)
21. von Wright, G.H.: *Norm and Action*. Routledge and Kegan Paul, London (1963)
22. Kantabutra, S.: What do we know about vision? *Sustain. Lens. Sustain.* **15**, 8403 (2023). <https://doi.org/10.3390/su15108403>

23. Larwood, L., Falbe, C.M., Kriger, M.P., Miesing, P.: Structure and meaning of organizational vision. *Acad. Manag. J.* **38**(3), 740–769 (1995)
24. Ortegon-Sanchez, A., Tyler, N.: A conceptual model for transformative urban planning. *Trans. Res. Procedia* **13**, 6–17 (2016). <https://doi.org/10.1016/j.trpro.2016.05.002>
25. Konrad, K., Böhle, K.: Socio-technical futures and the governance of innovation processes: an introduction to the special issue. *Futures* **109**, 101–107 (2019)
26. Mager, A., Katzenbach, C.: Future imaginaries in the making and governing of digital technology: multiple, contested, commodified. *New Media Soc.* **23**(2), 223–236 (2021). <https://doi.org/10.1177/1461444820929321>
27. Bareis, J., Katzenbach, C.: Talking AI into being: the narratives and imaginaries of national AI strategies and their performative politics. *Sci. Technol. Human Values* (2021)
28. Saariluoma, P., Cañas, J.J., Leikas, J.: Designing for Life - A human perspective on technology development. Palgrave MacMillan. (2016). <https://doi.org/10.1057/978-1-137-53047-9>
29. Osborne, S.P.: From public service-dominant logic to public service logic: are public service organizations capable of co-production and value co-creation? *Public Manag. Rev.* **20**(2), 225–231 (2017). <https://doi.org/10.1080/14719037.2017.1350461>
30. Lanne, M., Nieminen, M., Leikas, J.: Organisational tensions in introducing socially sustainable AI. *AI Soc.* (2025). <https://doi.org/10.1007/s00146-025-02293-y>
31. Zhang, Y., Wildemuth, B.M.: Qualitative analysis of content. In: Wildemuth, B.M. (ed.) *Applications of Social Research Methods to Questions in Information and Library Science*, pp. 318–329. Bloomsbury Publishing, Second Edition (2017)
32. Sheila, J., Sang-Hyun, K.: *Dreamscapes of modernity: sociotechnical imaginaries and the fabrication of power*. Chicago, University of Chicago Press (2015). <https://doi.org/10.7208/9780226276663>
33. Chalmers, D.J.: What is conceptual engineering and what should it be? *Inquiry* (2020). <https://doi.org/10.1080/0020174X.2020.1817141>



Security and Ethics in the Use of Computing Technologies and the Internet

Laercio Cruvinel Júnior^{1,2}(✉) , António Cabeças¹ ,
and Adriana Lopes Fernandes³

¹ Universidade Autónoma de Lisboa Luís de Camões, 1169-023 Lisboa, Portugal
{lcruvinel, acabecas}@autonoma.pt

² Autónoma TechLab, 1169-023 Lisboa, Portugal

³ Escola Superior Náutica Infante Dom Henrique, 2770-058 Paço de Arcos, Portugal
adrianafernandes@enautica.pt

Abstract. The rapid adoption of computing technologies and the Internet has transformed various aspects of society, including education, communication, commerce, and governance. While computer and communication technologies offer significant benefits, they also present complex ethical and security challenges. This paper explores the ethical and security dimensions of computing technologies, focusing on issues such as data privacy, algorithmic bias, cybersecurity threats, and digital well-being. Through a detailed analysis of these challenges, the paper examines how data collection, automated decision-making, and digital surveillance can undermine user autonomy, exacerbate inequalities, and compromise user privacy. The discussion is guided by ethical frameworks, including deontological and consequentialist perspectives, providing a balanced view of the ethical implications of technology use. The paper also proposes best practices for ethical technology integration, including clear data protection policies, bias mitigation strategies, transparent AI design, and user education programs. By promoting digital literacy and fostering a culture of ethical technology use, institutions can harness the benefits of computing technologies while minimizing risks. This paper emphasizes the need for a collaborative approach involving educators, administrators, developers, and policymakers to ensure that technology serves as a tool for empowerment rather than exploitation.

Keywords: Cybersecurity · Ethical AI · Digital Privacy · Algorithmic Bias · Digital Ethics · Regulatory Frameworks

1 Introduction

The rapid advancement of computing technologies and the pervasive use of the Internet have transformed modern society, reshaping education, business, communication, and governance [2]. In educational contexts, digital platforms provide students with personalized learning experiences and enable educators to automate administrative tasks [3]. In business, artificial intelligence (AI) and data analytics enhance decision-making and customer engagement [4]. Governments leverage digital tools for efficient service

delivery and public communication [5]. These technologies have become indispensable for economic growth, knowledge dissemination, and social interaction.

However, alongside these benefits, the widespread adoption of computing technologies has introduced a range of ethical and security challenges. Data privacy is a central concern, as users' personal information is often collected, stored, and processed without sufficient transparency or user consent [12]. Algorithmic bias can perpetuate social inequalities, with automated decision-making systems disproportionately affecting marginalized groups [13]. Cybersecurity threats, including data breaches, phishing, and ransomware attacks, have become more sophisticated, posing significant risks to individuals and organizations [27]. Digital burnout has emerged as a growing problem, with constant connectivity and digital monitoring contributing to stress and cognitive fatigue [23]. Finally, integrating artificial intelligence (AI) in decision-making raises ethical questions about user autonomy, transparency, and accountability.

Given the scope and complexity of these issues, this paper aims to provide a comprehensive analysis of the ethical and security challenges associated with computing technologies and the Internet. It explores how these technologies can impact user privacy, fairness, autonomy, and well-being, using ethical frameworks such as deontology [6, 7], consequentialism [14], and virtue ethics [15] to critically evaluate these concerns. The paper also proposes best practices for the responsible and ethical use of computing technologies, offering practical recommendations for educators, administrators, developers, and policymakers.

The remainder of this paper is structured as follows. Section 2 provides a review of the background on the theoretical foundations of the study, including ethics in computing and the related frameworks. Section 3 discusses the ethical and security challenges of computing technologies, including data privacy, algorithmic bias, autonomy, and cybersecurity. Section 4 explores specific security threats in computing, such as cybercrime, adversarial AI, and digital surveillance. Section 5 presents best practices for ethical technology use, including ethical implementation, AI literacy, human-AI collaboration, and responsible innovation. Finally, Sect. 6 offers conclusions and recaps the main insights of the work.

2 Background and Foundations

2.1 Ethics in Computing Technologies

The ethical use of computing technologies is a critical concern in modern digital societies. As computing systems become increasingly integrated into everyday life, their ethical implications extend far beyond technical functionality, touching on fundamental human rights, user autonomy, and social justice [1]. This section discusses four key ethical concepts—privacy, autonomy, fairness, and transparency—and their relevance to computing technologies.

Privacy

Privacy is a critical ethical concern in computing technologies, referring to an individual's right to control their personal information [11]. Privacy violations occur when user data is collected without informed consent, stored insecurely, or shared without user

knowledge. Social media platforms and educational tools, like learning management systems (LMS), often collect sensitive data with limited transparency. Protecting privacy requires strong data protection measures, such as encryption, secure authentication, and clear user information.

Autonomy

Autonomy is the ethical principle that individuals should have the freedom to make informed decisions without undue influence. In computing, autonomy is compromised when automated systems make decisions without user understanding or input, such as social media algorithms curating content or adaptive learning platforms dictating educational paths [13]. Preserving autonomy requires user-centered design, control options, and explainable algorithms.

Fairness

Fairness is an ethical principle that ensures individuals are treated equitably and without discrimination. In computing, fairness is critical in algorithm design, where biased training data can lead to discriminatory outcomes [13]. Examples include facial recognition errors affecting minority groups and biased automated grading. Ensuring fairness requires bias audits, diverse training data, and inclusive design practices.

Transparency

Transparency is the ethical principle that users should have clear, accessible information about how computing technologies work, including data collection, algorithms, and decision-making [9, 29]. Without transparency, “black box” systems can obscure decision logic, leading to mistrust.

Privacy, autonomy, fairness, and transparency are interconnected ethical principles in computing. Transparency can enhance autonomy by enabling informed decisions, while fairness relies on clear, privacy-protected algorithms. However, these principles can conflict—for example, full transparency may compromise privacy by exposing sensitive data. Balancing these principles requires a context-sensitive approach, making ethics a core aspect of technology design and governance.

2.2 Theoretical Foundations

Deontology

Immanuel Kant (1724–1804) was the author of several relevant works in many areas of philosophy, including ethics. Kant’s ethics, known as **deontology**, is centered on duties defined by what is right and wrong. In this ethical framework, **right actions are those we are morally required to perform**, while **wrong actions are those we are morally forbidden from doing**. Unlike value-based ethics, which focus on good and bad outcomes, deontology emphasizes that moral duties take precedence. Kant argues that all duties can be derived from a fundamental principle known as the **categorical imperative** [6]. **Deontological ethics** provides a framework for computing technologies that emphasizes user rights, informed consent, and fairness through respect for autonomy and privacy protection measures. Several studies address ethics in computing by applying deontological and rights-based perspectives to user rights, informed consent, and fairness. For example, [7–9] apply deontological principles—such as the generalization principle and respect for autonomy—to issues in natural language processing,

cybersecurity, and systems development. Other studies use rights-based approaches that align with deontological ethics and emphasize informed consent as central to protecting personal autonomy and privacy [10–12]. Framework proposals for algorithmic fairness [13] and digital human rights [12] demonstrate how ethical design can address inequities in information access and security.

Consequentialism and Virtue Ethics

Consequentialism is an ethical theory asserting that the moral value of an action is determined solely by its outcomes. In other words, actions are considered right if they lead to the best possible consequences and wrong if they result in harmful outcomes. This perspective is rooted in the intuition that what truly matters is the future—since the past is unchangeable, ethical decision-making should focus on creating the most favorable results moving forward [14].

Virtue ethics is a normative ethical approach that emphasizes moral character and virtues, such as kindness, honesty, and courage, rather than rules (deontology) or outcomes (consequentialism). It defines ethical behavior by the individual's character. For example, a virtue ethicist would see helping someone as an act of compassion, reflecting good character [15].

The integration of consequentialist evaluations and virtue-based approaches helps institutions and developers make ethical decisions about security measures while considering both practical benefits and moral risks. Studies invoking consequentialism—often through utilitarian ethics—argue for risk–benefit analysis and proportionality assessments to balance security gains against ethical risks such as privacy infringements, surveillance, and profiling [10, 16]. Papers applying virtue ethics emphasize that the cultivation of character traits like courage, temperance, and open-mindedness can guide ethical decision-making in cybersecurity. For example, Nakato et al. highlight the role of virtues in shaping cybersecurity behavior in banking [17], while Barford and Pieters and Consoli contend that ethical training and the integration of virtues into design practices are essential for responsible technology development [18, 19]. The findings indicate that integrating consequentialist evaluations with virtue-ethics principles may help developers and institutions balance the benefits of computing technologies against their ethical risks.

3 Ethical Challenges in Computing Technologies

3.1 Data Privacy and Surveillance

Privacy is one of the most pressing ethical issues in computing technologies [20]. As digital platforms become essential in daily life, the routine collection, storage, and processing of personal data raise ethical concerns about user autonomy, consent, and control. Users often share data without understanding how it will be used or who will access it [24]. Ensuring ethical data practices requires transparency, user consent, secure storage, and data minimization. Compliance with legal frameworks, such as GDPR, further protects user rights.

Ultimately, protecting user privacy in computing technologies is not just a technical challenge but an ethical imperative. It requires a commitment to respecting user autonomy, maintaining clear communication, and continuously updating privacy practices in response to emerging threats and user expectations [21].

3.2 Algorithmic Bias and Fairness

Artificial intelligence (AI) and machine learning are widely used in sectors like health-care, finance, and education, offering efficiency and personalized recommendations. However, they are prone to algorithmic bias, leading to unfair outcomes, especially for marginalized groups. Bias can stem from three main sources: biased training data, flawed model design, and improper deployment.

Addressing these ethical challenges requires a nuanced understanding of fairness, which can be interpreted through multiple ethical lenses:

Deontological Perspective (Duty-Based Ethics)

AI systems must treat all individuals with equal respect and dignity, making fairness a core design principle. This means that factors like gender, race, or age should be excluded from decision-making processes. For example, an AI hiring system should evaluate candidates solely on their qualifications.

Consequentialist Perspective (Outcome-Based Ethics)

In contrast, a consequentialist approach to AI focuses on outcomes, justifying algorithmic trade-offs if they maximize overall benefits, such as efficiency or accuracy. For instance, a predictive policing algorithm may target high-crime areas for resource allocation. However, this approach can normalize systemic inequalities, disproportionately affecting marginalized communities.

Virtue Ethics Perspective (Character-Based Ethics)

A virtue ethics approach emphasizes the moral character of those designing and deploying AI, promoting values like fairness, honesty, and empathy. Developers should consider diverse perspectives in model design, while educators should teach critical AI literacy, enabling users to recognize and challenge biased outcomes.

3.3 Autonomy, Digital Ethics, and Work-Life Balance

AI-driven recommendations and automated decisions can undermine user autonomy, especially when they function as “black boxes” with opaque decision-making [20]. This lack of transparency restricts users’ ability to make informed choices, violating their moral agency. A deontological view sees this as ethically problematic, as users deserve to understand and control how AI affects them. Conversely, a utilitarian view may justify automation for efficiency, but this must be balanced against the psychological and social costs of reduced user control.

3.4 Cybersecurity and Digital Exhaustion

Cybersecurity is essential for protecting digital systems, networks, and data from unauthorized access and damage. However, excessive cybersecurity measures can unintentionally harm user well-being, undermining autonomy and causing stress [22]. Balancing security with user mental health is a critical ethical challenge.

The Ethical Imperative of Cybersecurity

Organizations, especially those handling sensitive data like educational, healthcare, and financial institutions, have an ethical duty to ensure cybersecurity. Essential measures include data encryption, secure authentication, access control, and regular vulnerability assessments [25]. From a deontological perspective, organizations are morally obligated to protect user data, and failure to do so compromises user rights.

The Paradox of Digital Exhaustion

Sophisticated cybersecurity measures can lead to digital exhaustion, where users feel overwhelmed by constant monitoring and security checks [23]. In educational settings, learning management systems track student behavior, while workplaces use productivity tracking and biometric authentication. Although these measures enhance security, they can create anxiety and mistrust, highlighting the ethical tension between protection and user well-being.

Balancing Security and Well-Being – The Need for a Human-Centric Approach

The ethical tension between cybersecurity and user well-being can be resolved with a balanced approach that protects digital systems while respecting user autonomy and mental health. Organizations should prioritize user empowerment, transparency, and well-being, creating a secure environment without excessive surveillance.

4 Security Threats in Computing: Cybercrime, Adversarial AI, and Digital Surveillance

Cybersecurity is crucial as digital technologies face threats like data breaches, ransomware, and identity theft [26]. Organizations use advanced measures like encryption, multi-factor authentication, and network monitoring to protect systems. However, these measures create ethical trade-offs, such as privacy concerns with continuous monitoring and accessibility challenges with complex authentication methods.

4.1 Adversarial AI in Cybersecurity: Benefits and Risks

AI has transformed cybersecurity by enabling rapid threat detection, predictive analysis and automated responses. AI-powered systems can identify suspicious activities and prevent attacks [27]. But they also present risks, such as adversarial AI attacks where malicious actors manipulate models [28], and the “black box” effect, where AI decisions lack transparency, raising concerns about accountability and user control.

4.2 Digital Surveillance: Balancing Security and Privacy

Organizations use digital surveillance technologies, such as network monitoring, biometric authentication, and facial recognition, to enhance security. While these tools help detect threats and verify identities, they raise ethical concerns. Surveillance can infringe on privacy, reduce free expression, and create fear in workplaces or self-censorship in educational settings. Balancing security and user rights is crucial.

From a deontological perspective, pervasive surveillance violates individuals' rights by treating them as means to institutional efficiency rather than as ends in themselves. Users deserve transparency regarding data collection, monitoring, and control. Excessive surveillance undermines privacy, discourages free thought, and stifles critical thinking and innovation [21].

5 Proposed Solutions and Best Practices

Ensuring the ethical, secure use of computing technologies requires a multi-faceted approach, combining technical measures, regulatory policies, education. While best practices can promote responsible technology use, their implementation faces challenges such as technical limitations, resource constraints, and stakeholder resistance.

5.1 Ethical AI Development

The following best practices are essential for ethical AI development and adherence to ethical principles that prioritize human rights, fairness, and accountability:

- **Transparency:** AI systems must provide clear, accessible information about how they function, how decisions are made, and what data they use, fostering user trust and accountability. For example, an AI-powered recruitment system should disclose how it assesses candidate qualifications.
- **Accountability:** Developers must implement mechanisms for detecting, correcting, and preventing biases in AI models, such as regular bias audits and explainable AI (XAI) techniques [29]. This ensures that AI decisions, like those in financial credit scoring, can be reviewed and corrected if biased.
- **Inclusivity:** AI models should be trained on diverse, representative datasets that reflect various demographic groups to avoid perpetuating social inequalities. For instance, facial recognition algorithms must be tested across different populations to prevent racial or gender bias.
- **Human-Centered Design:** AI applications should prioritize user well-being, offering user control options, clear explanations for recommendations, and features to minimize digital exhaustion. For example, educational AI tools should allow students to customize their learning paths rather than rigidly following algorithmic suggestions.

Challenges for implementation include **Complexity and Cost**, **Resistance from Developers**, and **Bias in Training Data**.

5.2 Strengthening Cybersecurity Measures

Organizations and individuals should adopt the following best practices for protecting user data, maintaining system integrity, and preventing unauthorized access:

- **Encryption and Secure Authentication:** Sensitive data should be protected with strong encryption methods, and multi-factor authentication (MFA) should be implemented to ensure secure access to systems.
- **Regular Security Audits:** Organizations should conduct periodic vulnerability assessments and penetration testing to identify and address security weaknesses.
- **User Awareness Programs:** Educational initiatives should teach users about common cybersecurity threats, such as phishing attacks, social engineering, and malware.

Challenges for implementation include **Evolving Threat Landscape**, **Resource Constraints**, and **Difficult User Compliance**.

5.3 Regulatory and Policy Frameworks

Governments and regulatory bodies play a crucial role in ensuring the ethical and secure use of computing technologies. Effective regulatory measures should include:

- **Data Protection Laws:** Governments should enforce compliance with global data protection standards, such as the General Data Protection Regulation (GDPR) in Europe or the California Consumer Privacy Act (CCPA) in the United States.
- **AI Ethics Guidelines:** Regulatory bodies should establish clear ethical standards for AI development and deployment, including guidelines for transparency, accountability, and bias mitigation.
- **Consumer Rights Protection:** Organizations must be required to maintain transparent data collection practices, ensure user consent, and provide mechanisms for users to challenge or delete their data.

Challenges for implementation include **Regulatory Compliance Costs**, **Jurisdictional Conflicts**, and **Regulatory Lag** (due to rapid technological advancements).

5.4 Digital Literacy and User Responsibility

Digital literacy is essential for empowering individuals to navigate the ethical and security challenges of digital environments. Educational programs should focus on the following areas:

- **Privacy Awareness and Critical Thinking:** Users should be taught how to safeguard their personal data, including understanding data privacy settings, recognizing phishing attempts, managing online permissions, critically evaluate online information, recognize misinformation, and question algorithmic recommendations.
- **Ethical Technology Use:** Educational programs should promote ethical behavior in digital environments, encouraging responsible AI use, respectful communication, and awareness of digital well-being.

Challenges for implementation include **Accessibility**, **Rapid Technological Change**, and **Lack of User Engagement**.

5.5 Towards a Comprehensive Ethical Framework

To ensure that these best practices are effectively implemented, organizations must adopt a comprehensive ethical framework that integrates technical, regulatory, and educational measures. This framework should include:

- **Regular Ethical Assessments:** Organizations should conduct periodic reviews of their AI and cybersecurity practices, identifying ethical risks and areas for improvement.
- **Collaborative Governance, Empowerment and Inclusion:** Institutions should involve diverse stakeholders—students, educators, developers, policymakers, and users—in the development of ethical policies. Digital literacy programs should prioritize accessibility and inclusivity, ensuring that all users can benefit from ethical technology use.
- **Adaptive Policies:** Ethical guidelines should be regularly updated to address emerging ethical challenges and technological advancements.

By adopting these best practices, organizations can create secure, transparent, and ethically responsible digital environments that protect user rights while promoting innovation.

6 Conclusions

The rapid advancement of computing technologies and the Internet has transformed society, enhancing education, communication, business, and governance. However, it also brings ethical and security challenges, such as data privacy risks, algorithmic bias, cybersecurity threats, digital burnout, and AI ethics. This paper critically analyzes these issues and proposes best practices for responsible technology use.

6.1 Key Insights

This paper has highlighted the ethical and security challenges associated with computing technologies, focusing on four primary areas:

- **Ethical Challenges:** Issues such as data privacy, algorithmic bias, user autonomy, and digital well-being are critical concerns in the digital age. Misuse of personal data, lack of transparency in algorithmic decision-making, and constant pressure of digital engagement can undermine user rights and mental health.
- **Security Risks:** Cybercrime, adversarial AI, and digital surveillance pose significant threats to individuals and organizations. While advanced security measures, such as end-to-end encryption and AI-powered threat detection, can protect digital systems, they also raise ethical concerns, such as excessive monitoring and user manipulation.
- **Best Practices for Ethical Technology Use:** This paper proposed a multi-faceted approach to promoting ethical and secure technology use, including:
 - **Ethical AI Development:** Ensuring transparency, accountability, inclusivity, and human-centered design.

- **Strengthening Cybersecurity:** Implementing robust encryption, conducting regular security audits, and promoting user awareness.
- **Regulatory and Policy Frameworks:** Enforcing data protection laws, establishing AI ethics guidelines, and protecting consumer rights.
- **Digital Literacy and User Responsibility:** Promoting privacy awareness, critical thinking, and ethical technology use.
- **The Role of Stakeholders:** Effective ethical governance of computing technologies requires collaboration among students, educators, administrators, policymakers, and developers. Each stakeholder has a unique role in promoting responsible technology use, from educating users to designing secure systems and establishing ethical policies.

6.2 Final Considerations

The ethical and secure use of computing technologies is not just a technical challenge—it is a fundamental responsibility that requires a commitment to human rights, user autonomy, and social justice. In an increasingly digital world, transparency, fairness, and accountability must be at the heart of technology design, deployment, and governance. Organizations must prioritize user empowerment, providing clear explanations for algorithmic decisions, respecting user privacy, and ensuring that digital systems enhance well-being rather than undermining it.

At the same time, security measures must be carefully balanced against user rights. Robust cybersecurity practices are essential for protecting user data and maintaining system integrity, but they must not become tools for excessive surveillance or control. Ethical AI design should focus on reducing biases, promoting fairness, and maintaining user agency.

Ultimately, responsible technology use depends on a culture of ethical reflexivity—where institutions continuously evaluate and improve their technology practices, adapting to emerging challenges while maintaining a commitment to ethical principles. This reflexivity is not only a matter of compliance but a fundamental aspect of maintaining user trust and fostering a resilient digital society.

References

1. Stahl, B.C., Timmermans, J., Mittelstadt, B.D.: The Ethics of Computing: A Survey of the Computing-Oriented Literature. *ACM Comput. Surv.* **48**(4), 1–38 (2016). <https://doi.org/10.1145/2871196>
2. Yamin, M.: Information technologies of 21st century and their impact on the society. *Int. J. Inform. Technol.* **11**(4), 759–766 (2019). <https://doi.org/10.1007/s41870-019-00355-1>
3. Timotheou, S., et al.: Impacts of digital technologies on education and factors influencing schools' digital capacity and transformation: a literature review. *Educ. Inf. Technol.* **28**(6), 6695–6726 (2023). <https://doi.org/10.1007/s10639-022-11431-8>
4. Rimón, S.M.T.H.: Leveraging artificial intelligence in business analytics for informed strategic decision-making: enhancing operational efficiency, market insights, and competitive advantage. *J. Artif. Intell. General Sci.* **6**(1), 600–624 (2024). <https://doi.org/10.60087/jaigs.v6i1.278>

5. Falco, E., Kleinhans, R.: Beyond technology: Identifying local government challenges for using digital platforms for citizen engagement. *Int. J. Inform. Manage.* **40**, 17–20 (2018). <https://doi.org/10.1016/j.ijinfomgt.2018.01.007>
6. Johnson, R., Cureton, A.: Kant's Moral Philosophy. *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), Edward, N.Z., Nodelman, U.(eds.), <https://plato.stanford.edu/archives/fall2024/entries/kant-moral/> (2024)
7. Prabhumoye, S., Boldt, B., Salakhutdinov, R., Black, A. W.: Case study: deontological ethics in NLP. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3784–3798, Online. Association for Computational Linguistics (2021). <https://doi.org/10.18653/v1/2021.naacl-main.297>
8. Posey, C., Folger, R.: An exploratory examination of organizational insiders' descriptive and normative perceptions of cyber-relevant rights and responsibilities. *Comput. Secur.* **99**, 102038 (2020). <https://doi.org/10.1016/j.cose.2020.102038>
9. Rahanu, H., Georgiadou, E., Siakas, K., Ross, M.: Health, safety, and well-being through ethical behaviours in computer and information systems development and deployment. *J. Softw. Evol. Proc.* **32**, e2225 (2020). [https://doi.org/10.1002/smr.2225\(2020\)](https://doi.org/10.1002/smr.2225(2020))
10. Loi, M., Christen, M.: Ethical frameworks for cybersecurity. In: Christen, M., Gordijn, B., Loi, M. (eds.) *The Ethics of Cybersecurity*, pp. 73–95. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-29053-5_4
11. Walters, G.J.: Privacy and security: an ethical analysis. *SIGCAS Comput. Soc.* **31**(2), 8–23 (2001). [https://doi.org/10.1145/503345.503347\(2001\)](https://doi.org/10.1145/503345.503347(2001))
12. Tsiakis, T.: The role of information security and cryptography in digital democracy: (human) rights and freedom. In: *Information Resources Management Association (Ed.), Human Rights and Ethics: Concepts, Methodologies, Tools, and Applications*, pp. 1564–1580. IGI Global Scientific Publishing (2015). <https://doi.org/10.4018/978-1-4666-6433-3.ch086>
13. Udoh, E.S., Yuan, X., Rorissa, A.: A framework for defining algorithmic fairness in the context of information access. *Proc. Assoc. Inform. Sci. Technol.* **61**(1), 667–672 (2024). <https://doi.org/10.1002/pr2.1077>
14. Sinnott-Armstrong, W.: Consequentialism. *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), Zalta, E.N., Nodelman, U.(eds.) (2023). <https://plato.stanford.edu/archives/win2023/entries/consequentialism/>
15. Hursthouse, R., Pettigrove, G.: Virtue Ethics. *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), Zalta, E.N., Nodelman, U.(eds.) (2023). <https://plato.stanford.edu/archives/fal12023/entries/ethics-virtue/>
16. Smith, A.D.: E-security issues and policy development in an information-sharing and networked environment. *Aslib Proc.* **56**(5), 272–285 (2004). <https://doi.org/10.1108/00012530410560878>
17. Nakato, R., Kaggwa, F.: Establishing the influences of cardinal virtues on employees' cyber security ethical behavior in the banking sector in Uganda. *Europ. J. Technol.* **6**, 1 – 1–13 (2022)
18. Barford, L.: Contemporary virtue ethics and the engineers of autonomous systems. In: *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pp. 1–7. Medford, MA, USA (2019). <https://doi.org/10.1109/ISTAS48451.2019.8937855> (2019)
19. Pieters, W., Consoli, L.: Vulnerabilities and responsibilities: dealing with monsters in computer security. *J. Inform. Commun. Ethics Soc.* **7**(4), 243–257 (2009). <https://doi.org/10.1108/14779960911004499>
20. Zostant, M., Chataut, R.: Privacy in computer ethics: Navigating the digital age. *Comput. Sci. Inf. Technol.* **4**(2), 183–190 (2023). <https://doi.org/10.11591/csit.v4i2.p183-190>
21. van Baalen, S.: 'Google wants to know your location': The ethical challenges of fieldwork in the digital age. *Res. Ethics* **14**(4), 1–17 (2018). <https://doi.org/10.1177/1747016117750312>

22. Jawad, L.A.: Security and privacy in digital healthcare systems: challenges and mitigation strategies. *Abhigyan* **42**(1), 23–31 (2024). <https://doi.org/10.1177/09702385241233073>
23. Pinto, M.I.F.: Devem as instituições de ensino superior estar preocupadas com o digital burnout? Uma perspectiva sobre a satisfação no trabalho. Master thesis (2023). <http://hdl.handle.net/10400.5/30117>
24. Knight, T.D.: Genealogical ethics in the united states and the popularization of genealogical research in the digital age. *Genealogy* **8**(3), 78 (2024). <https://doi.org/10.3390/genealogy8030078>
25. Faltermaier, S., Strunk, K.S., Obermeier, M., Fiedler, M.: Managing organizational cyber security - the distinct role of internalized responsibility. In: Hawaii International Conference on System Sciences (2023)
26. Shillair, R., Cotten, S.R., Tsai, H.-Y.S., Alhabash, S., LaRose, R., Rifon, N.J.: Online safety begins with you and me: convincing internet users to protect themselves. *Comput. Human Behav.* **48**, 199–207 (2015). <https://doi.org/10.1016/j.chb.2015.01.046>
27. Wang, R.: AI use in enhancing cybersecurity for safeguarding digital information. In: 2023 International Conference on Applied Physics and Computing (ICAPC), Applied Physics and Computing (ICAPC), 2023 International Conference on, ICAPC, pp. 412–419 (2023). <https://doi.org/10.1109/ICAPC61546.2023.00082>
28. Malatji, M., Tolah, A.: Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. *AI Ethics* **5**(2), 883–910 (2025). <https://doi.org/10.1007/s43681-024-00427-4>
29. Kyrimi, E., McLachlan, S., Wohlgemut, J.M., et al.: Explainable AI: definition and attributes of a good explanation for health AI. *AI Ethics* **5**(4), 3883–3896 (2025). <https://doi.org/10.1007/s43681-025-00668-x>



Social Risks of Brain-Machine Interface Usage: Questionnaire Survey of People with and without Disabilities

Yohko Orito¹ , Tomonori Yamamoto¹ , Shizuka Suzuki¹ , Hidenobu Sai¹ ,
Kiyoshi Murata² , Yasunori Fukuta² , Taichi Isobe³ , and Masashi Hori⁴ 

¹ Ehime University, Matsuyama 790-8577, Ehime, Japan
orito.yohko.mm@ehime-u.ac.jp

² Meiji University, Chiyoda 101-8301, Tokyo, Japan

³ Health Sciences University of Hokkaido, Ishikari-gun 061-0293, Hokkaido, Japan

⁴ Waseda University, Shinjuku 169-8050, Tokyo, Japan

Abstract. In recent years, brain-machine interfaces (BMIs) have been developed and have found practical applications, in particular, for people with disabilities. However, ethical concerns, such as the protection of mental privacy and other social issues related to such BMI usage remain underexplored and are often overlooked. To fill the gaps, a questionnaire survey of people with and without disabilities was conducted. This study attempts to identify the overall tendencies and differences between the responses of the two groups in terms of the usefulness, risks and ethical concerns of BMIs based on the survey results.

Keywords: Brain Machine Interface · Support for People with Disabilities · Cyborgisation · Questionnaire Survey

1 Introduction

In recent years, brain-machine interfaces (BMIs) or brain-computer interfaces (BCIs), which use brain signals to control digital devices, have been developed and have found practical applications. For instance, a system for reconstructing mental imagery from brain signal data has been explored [1]. Among these applications, the use of BMIs in the field of social welfare is expected to be significant, in particular, for people with disabilities who cannot move their limbs. These technologies have the potential to assist daily activities and facilitate communication for them. It could be observed that BMI devices were used to support people with amyotrophic lateral sclerosis (ALS) and brain stroke [2, 3]. However, ethical concerns, such as the protection of mental privacy and other social issues related to BMI usage, particularly among people with disabilities, remain underexplored and are often overlooked.

To address these gaps, the authors conducted experiments using BMI systems and semi-structured interviews to examine the ethical and social issues related to the use of BMIs for people with disabilities. In these experimental surveys, participants with and

without disabilities, as well as experts from the social welfare field, were asked to wear a headset-type non-invasive BMI device (an electroencephalography (EEG) headset) to remotely operate a robotic arm. Semi-structured interviews were conducted at three stages: before, during, and after the experiment. Interview questions were developed to investigate attitudes toward the usefulness and potential risks of BMIs. Based on prior survey results, some ethical and social issues associated with the use of BMI devices for people with disabilities have been identified (see, [4–7]).

On the other hand, such experiments and semi-structured interviews as qualitative surveys required to set up an experimental environment with a pre-programmed robotic arm and an EEG headset, and thus took a considerable amount of time. These requirements have inherent limitations, including the inability to conduct research on a large number of people with different disabilities as participants. Not only experimental surveys using the BMI system in practice and qualitative surveys such as semi-structured interviews with participants, but also quantitative surveys on the usefulness and ethical issues of the BMI system are necessary to collect the opinions of a larger number of people with and without disabilities.

We therefore conducted a questionnaire survey on general attitudes and opinions regarding the potential usefulness of and ethical concerns about cyborg technologies including BMIs. It was also intended to enable a comparison of opinions and recognitions between people with and without disabilities. Based on the survey results, this study attempts to identify the overall tendencies and differences between the responses of the two groups in terms of the usefulness, risks and ethical concerns of BMIs.

2 Overview of the Questionnaire Survey

2.1 Attribution of the Respondents: People with and Without Disabilities

In this study, a questionnaire survey with two groups (people with and without disabilities) was administered with the same question items as much as possible in Japan. The questionnaire survey for both groups comprised a fact sheet; questions about opinions on social welfare support for people with disabilities; recognition of BMIs or BCIs; usefulness and ethical concerns regarding BMI devices; and questions on whether BMIs contribute to or detract from individual life, workplace and society.

For people without disabilities, an online survey was conducted using Google Forms from February to March 2024. The respondents were recruited by a survey research company, so that the distribution of gender, age, and place of residence (prefecture) was almost uniform, respectively. Of the 848 responses, 683 were deemed valid. The gender and age distributions of the respondents are listed in Table 1. Among them, 264 respondents (38.65%) were full-time employees, 125 respondents (18.30%) were students, 112 respondents (16.40%) were non-regular employees, 94 (13.76%) were housewives/househusbands, 51 respondents (7.47%) were unemployed, and other respondents included self-employed, and part-time workers.

For people with disabilities, an online survey using Google Forms and a written offline survey were conducted from March to November 2024, targeting only those with physical disabilities. Of the 89 responses, 87 were deemed valid by the authors. The attributes of the survey participants are listed in Tables 2 and 3. The assistive devices the

respondents with disabilities used ('What kind of assistive devices do you currently use?', multiple answers were allowed) were as follows: "Wheelchairs" (85, 97.70%), "Electric wheelchairs" (60, 68.97%), "Orthotic device" (10, 11.49%), and "Seat holding devices" (10, 11.49%). Just a small number of respondents used "Eyewear", "Hearing aids", "Walking aids", "Communicators for people with severe disabilities", "Prosthetic limb", and "Safety rod for blind people". As summarised in the tables, a slight demographic bias in terms of gender and age was observed, in contrast to the survey of people without disabilities. Notably, in Japan, approximately 9.2% of the population has some form of disabilities [8]. Therefore, although the number of responses from people with disabilities in this survey was small, 87 responses out of 770 (11.30%) are considered to be a sufficient ratio.

Table 1. Gender and age distributions: People without disabilities $n = 683$

		Age						Total
		10s	20s	30s	40s	50s	60s (60–65)	
Gender	Male	57	64	55	57	50	54	337
	Female	57	58	55	53	57	63	343
	prefer not to answer	1	0	1	1	0	0	3
	Total	115	122	111	111	107	117	683

Table 2. Gender and age distributions: People with disabilities $n = 87$

		Age						Total
		10s	20s	30s	40s	50s	60s	
Gender	Male	1	7	12	16	16	7	59
	Female	0	6	10	7	1	3	27
	prefer not to answer	0	0	0	1	0	0	1
	Total	1	13	22	24	17	10	87

Table 3. Types and symptoms of disabilities: People with disabilities (multiple answers were allowed in terms of symptoms) $n = 87$

SYMPTOMS (Multiple Answer)	TYPES				
	Congenital disabilities	Acquired disabilities	Both (congenital/acquired)	No answer	Total
Physical limb disability	39	43	1	1	84

(continued)

Table 3. (continued)

SYMPTOMS (Multiple Answer)	TYPES				
	Congenital disabilities	Acquired disabilities	Both (congenital/ acquired)	No answer	Total
Voice, language and mastication disability	4	6	0	0	10
Internal disability due to internal organ function and other diseases	0	6	0	0	6
Blindness and visual disability	1	1	0	0	2
Deafness and disequilibrium disability	0	1	0	0	1
Other	2	1	1	0	4

3 Result of Questionnaire Survey

3.1 Recognition on BMIs or BCIs

As shown in Table 4, more than 70% of the respondents from the both respondent groups (557/770) were not aware of BMIs or BCIs at all, and only 89 respondents answered, “I know a lot about BMIs or BCIs” or “I have heard of the term BMI or BCI and know to some extent what kind of device it is”. Among them, 14 respondents (15.73%) had used the devices, whereas 40 respondents (44.94%) answered “I have seen it, but never used it”. According to these, BMIs or BCIs were not popular devices for either group of respondents, and in fact no statistically significant differences were observed between the two groups regarding the answers to question on recognitions on BMIs or BCIs ($p = .5260$)¹.

¹ Pearson’s chi-squared test and Fisher’s exact test were applied depending on the numbers. The same applies hereafter.

Table 4. Recognitions on BMIs or BCIs $n = 770$

Do you know the term brain-machine interfaces (BMIs) or brain-computer interfaces (BCIs)?	People without disabilities		People with disabilities		Total
	Number	%	Number	%	
I know a lot about BMIs or BCIs	16	2.34%	1	1.15%	17
I have heard of the term BMIs or BCIs and know to some extent what kind of device it is	61	8.93%	11	12.64%	72
I have heard of the term BMIs or BCIs, but do not know what kind of device it is	107	15.67%	17	19.54%	124
I do not know at all	499	73.06%	58	66.67%	557
Total	683		87		770

In contrast, regarding the recognition of the expected benefits, possibilities and risks of BMIs, some significant differences were observed in responses between people with and without disabilities, as described in the sections below.

3.2 Expected Benefits and Possibilities of BMI Usage

The perceptions of the expected benefits and potential applications of BMIs by respondents with and without disabilities were explored through the question², ‘In what situations would you like to use BMI in your everyday life or work?’ The responses are presented in Table 5. Although the most expected use of BMI is to operate electronic devices for both groups, more respondents with disabilities indicated a desire to utilise BMI devices in their lives in a greater variety of possible ways. In fact, significant differences were observed between the two groups at the 1% level for the following options: “Operation of electronic devices”, “Leisure, hobbies”, “Assistance with daily activities in the house”, “Mobility assistance for going out and using public transport”, and “Creative activities”.

² Before being asked series of questions, all respondents were given a common explanation of what the BMIs are.

Table 5. Expected ways or purposes to use of BMIs (multiple answers were allowed) $n = 770$

	People without disabilities (683)		People with disabilities (87)		p -value (** $p < .01$)	
	Number	%	Number	%		
Operation of electronic devices (e.g. computers, smartphones, gaming devices)	203	29.72%	56	64.37%	.0000	**
Emergency calls (emergency communication)	198	28.99%	26	29.89%	.8625	
Identity authentication functions (as IC cards or personal authentication devices)	160	23.43%	24	27.59%	.3915	
Leisure, hobbies	142	20.79%	38	43.68%	.0000	**
Assistance with daily activities in the house, such as cooking, cleaning, toileting, and bathing	134	19.62%	40	45.98%	.0000	**
Mobility assistance for going out and using public transport	111	16.25%	44	50.57%	.0000	**
Communication with friends and acquaintances	121	17.72%	14	16.09%	.7075	
Creative activities such as writing and drawing	85	12.45%	23	26.44%	.0004	**
None, I do not know	47	6.88%	2	2.30%	.1568	
Other	5	0.73%	1	1.15%	-	

The question was followed by an optional open-ended item prompting the respondents ‘Please indicate the specific purposes for which you would like to use BMIs’. Consequently, 274 valid free descriptive responses from respondents without disabilities and 36 from those with disabilities were obtained. Table 6 presents the free descriptive answers from respondents without disabilities. A majority of them emphasised practical conveniences and efficiency, envisioning the use of BMIs for emergency contact and various daily tasks. This may highlight the desire for BMI to streamline life, increase personal safety, and extend human capabilities in various contexts.

Table 6. Categories and typical responses to expectations on the use of BMIs from respondents without disabilities (Optional, Free descriptive responses)

Category	Typical responses
Convenience and efficiency	'I would like to use it (BMIs) multitask while my hands are occupied with chores, such as cooking, cleaning, and other household chores.'
Assistance with daily activities	'I appreciate the potential usefulness of BMIs in daily activities, especially as my movements are becoming slower with age.'
Creative activities, leisure, hobbies	'Even if I cannot draw well by myself, I may draw better with BMIs.'
Emergency and safety	'In the event of an earthquake or other disaster, I think BMIs can be useful for contacting people without speaking out, especially in situations when I am trapped, such as under rubble, and need to conserve my energy.'
Operation of electronic devices	'It makes it possible to enter text on the smartphone just by thinking about it, and it is also easy to operate.'
Healthcare	'It is useful when receiving medical advice at a hospital.'
Communication	'It may be useful as part of a communication tool with overseas friends.'
Identity authentication functions	'It is good that it minimises the stress of identity authentication, enabling anyone to verify their identification safely.'

On the other hand, Table 7 presents the free descriptive responses provided by the respondents with disabilities to the same question. Several respondents expressed a desire to overcome the physical limitations and difficulties caused by their disabilities through the cyborg technology. This includes activities that lead to a more independent life and enhance social activities and communication. There was also a high expectation of improved quality of life, with some respondents wishing to use technology not only for functional support but also for hobbies and creative activities to enrich their lives.

Table 7. Categories and typical responses to expectations on the use of BMIs from respondents with disabilities (Optional, Free descriptive responses)

Category	Typical responses
Operation of electronic devices	'I have a cervical spine injury. Although I can usually operate a personal computer and smartphone. It would be nice if I could do it without using my hands, at least when texting or emailing. I would also like to operate TVs and DVD players with my brain signal!'

(continued)

Table 7. (continued)

Category	Typical responses
Assistance with activities of daily living	'As I cannot do things like bathing, toileting, housework and going out by myself at all, it would be easy if I could rely on a robot instead of a human caregiver someday. Unlike humans, I do not have to take care of robot caregivers. Further, due to my progressive disease, I may not be able to operate electronic devices in the future. I would like to be assisted by a robot at that time.'
Mobility assistance	'I would like to operate my electric wheelchair with brain signals.'
Communication	'If my current amyotrophic lateral sclerosis symptoms progress from quadriplegia Total Locked-in Syndrome, I will no longer be able to communicate with others.'
Enhancement and reinforcement of body functions	'My right arm cannot be used with a myoelectric prosthesis because the nerve is severed at the base of the neck. Therefore, I would like an arm that can be moved by brain signals. Further, for both lower limbs, I want an assistive device that will allow me to stand up and walk down steps. If that can be done (with the use of BMIs), I will not be afraid to cut off and replace my natural limbs.'
Creative activities	'I would like have something like a sub-arm when making and playing with plastic models.'
Privacy protection	'I cannot operate my mobile phone or PC by myself when I am in a wheelchair, so I would like to use it (BMIs) for such purposes. Although I have a caregiver, I do not want to be seen by them in some private situations.'

Moreover, their responses to the question, 'Regardless of your own daily life and the tasks or work you are engaged in on a daily basis, which field would you expect or desire BMIs to be applied to?' are listed in Table 8. Overall, similar to the results summarised in Table 5, the respondents with disabilities were more likely to use BMIs in various fields. Not only they have high expectations that they would be applied in the health care and social welfare fields, but more than half of them also expected that it would be applied for entertainment and education. Significant differences were observed between both groups at the 1% level for some options: "Social welfare", "Retail and marketing", "Sports", "Gaming, entertainment", "Learning, education", and "Travel, tourism".

Table 8. Expected applications of BMIs (multiple answers allowed) $n = 770$

	People without disabilities (683)		People with disabilities (87)		<i>p</i> -value (** $p < .01$)	
	Number	%	Number	%		
Health care	444	65.01%	49	56.32%	.1119	
Social welfare	334	48.90%	72	82.76%	.0000	**
Retail and marketing	76	11.13%	24	27.59%	.0000	**
Sports	74	10.83%	28	32.18%	.0000	**
Gaming, entertainment	148	21.67%	49	56.32%	.0000	**
Learning, education	154	22.55%	44	50.57%	.0000	**
Travel, tourism	96	14.06%	35	40.23%	.0000	**
None, I do not know	30	4.39%	30	34.48%	.0000	**
Other	5	0.73%	0	0.00%	-	

3.3 Perceived Risks and Harms of BMI Usage

The perceptions of both groups of the problems or risks caused by the BMI usage and their specific concerns were explored through the question, ‘If you use BMIs in your own daily life and work, which problems do you have concerns about?’ Table 9 lists the responses. “Malfunctioning, disoperation of the BMIs” was the most selected option among both groups. However, in addition to the responses related to the expected use of BMI, as explained above, respondents with disabilities were likely to indicate more comprehensive concerns about the problems associated with BMI usage. In particular, they expressed more concerns about issues such as “Malfunctioning of the BMIs”, “Privacy invasion: inappropriate use of personal data collected via BMIs”, and “Loss of electric power to BMI, no operation of it”. Significant differences existed among these options at the 1% level. Although a statistically significant difference was not found, the respondents without disabilities expressed slightly higher concerns about “Feeling like I am being treated like an object” than those with disabilities.

Table 9. Assumed problems, risks associated with BMI usage in daily life and work (Multiple answers were allowed) $n = 770$

	People without disabilities (683)		People with disabilities (87)		<i>p</i> -value (** $p < .01$)	
	Number	%	Number	%		
Malfunctioning, disoperation of the BMIs	374	54.76%	70	80.46%	.0000	**
Privacy invasion: inappropriate use of personal data collected via BMIs	165	24.16%	35	40.23%	.0013	**
Increased dependence on BMIs	162	23.72%	29	33.33%	.0505	
Loss of electric power to BMI, no operation of it	135	19.77%	45	51.72%	.0000	**
Increased feeling of social isolation	125	18.30%	19	21.84%	.4255	
Feeling like I am being treated like an object	99	14.49%	11	12.64%	.6421	
None, I do not know	32	4.69%	0	0.00%	.0756	
Other	2	0.29%	2	2.30%	-	

It should be noted that additional options were provided only for the respondents with disabilities because the following options were not appropriate for the respondents without disabilities; “Increased difficulties in providing care and communicating by human caregivers”, “The necessity for human caregivers decreases”, “Reduction in the number of welfare services available for people with disabilities”, “Decreased opportunities to go out”, and “Decreased need for human caregivers”. The responses to these additional options for people with disabilities are shown in Table 10.

Considering these additional results, the most sincere concern about BMI usage for the majority of the respondents with disabilities is “Malfunctioning, and disoperation of the BMIs”. Consequently, concerns about operational and security issues related to BMI usage were more pronounced among respondents with disabilities. Social and ethical risks and concerns regarding the use of BMI may be perceived as relatively low by both groups.

Table 10. Assumed problems, risks associated with BMI usage in daily life and work (Multiple answers were allowed; options applicable to only people with disabilities) $n = 87$

	Number	%
Increased difficulties in providing care and communicating by human caregivers	17	19.54%
Reduction in the number of welfare services available for people with disabilities	16	18.39%
Decreased opportunities to go out	12	13.79%
Decreased need for human caregivers	9	10.34%

3.4 Attitudes and Concerns Regarding Embedded BMIs

With regard to embedded BMIs, which may have more effective functions but also have more serious impacts on the users, the respondents were asked, ‘BMIs include a type of brain chip or embedded BMI, which is surgically embedded in the brain. Would you like to have an embedded BMI?’ The results (Table 11) reveal that more than half of the respondents did not like to use embedded BMIs, and no statistically significant differences were observed between the two groups regarding their responses to the question ($p = .2263$).

Table 11. Opinions on the use of embedded BMIs $n = 770$

	People without disabilities		People with disabilities		Total
	Number	%	Number	%	
Very willing to use embedded BMIs	26	3.81%	6	6.90%	32
Slightly willing to use embedded BMIs	40	5.86%	6	6.90%	46
Neither	182	26.65%	16	18.39%	198
Not much willing to use embedded BMIs	164	24.01%	27	31.03%	191
Never willing to use embedded BMIs	271	39.68%	32	36.78%	303
Total	683		87		770

The survey also explored the participants’ perception of the risks associated with embedded BMIs. Table 12 lists the responses to the question, ‘Which risks do you think would be posed by a BMI device, such as the brain chip, embedded in the brain?’ More than half of the respondents, regardless of disability status, selected “Concerns about the risk of physical or health conditions being worsened by the embedding operation”, indicating that the fears and concerns about the physical effects of embedded BMIs in

the body were relatively stronger than the other options. In contrast, it also appears that respondents with disabilities have more serious concerns and worries about the various risks posed by embedded BMIs than those without disabilities, as well as their general attitudes to the risks of using BMIs. In particular, significant differences exist between two groups about physical risks of embedded BMIs on their bodies at 1% level, more respondents with disabilities selected the “Concerns about the risk of physical or health conditions being worsened by the embedding operation” and “The inability to easily remove BMI devices”.

Table 12. Concerns and perceived risks associated with embedded BMIs (Multiple answers were allowed) *n* = 770

	People without disabilities (683)		People with disabilities (87)		<i>p</i> -value (** <i>p</i> < .01)	
	Number	%	Number	%		
Concerns about the risk of physical or health conditions being worsened by the embedding operation	406	59.44%	74	85.06%	.0000	**
The inability to easily remove BMI devices	337	49.34%	59	67.82%	.0012	**
Feeling like the machine is intrusive, as if something alien is invading the body	205	30.01%	31	35.63%	.2845	
The collection and storage of brain signal data	168	24.60%	27	31.03%	.1935	
Feeling of being controlled by the machine (BMI devices)	163	23.87%	21	24.14%	.9552	
Increased dependence on BMIs and the machines connected to bodies	137	20.06%	22	25.29%	.2565	
Concern about how others will perceive me	95	13.91%	15	17.24%	.4029	
I do not know	9	1.32%	0	0.00%	.5841	
others	11	1.61%	1	1.15%	-	

4 Discussions

Overall, the two groups showed similar tendencies in their knowledge of BMIs and BCIs, as well as in their experience of using BMI. In other words, almost all respondents, regardless of their disability status, did not demonstrate sufficient knowledge about BMIs and BCIs or experiences in operating BMI devices. They also tended to expect a range of benefits and possibilities and assumed that BMI devices would facilitate more efficient, safer, and more stimulating activities in their lives and society. In addition, most respondents in both groups exhibited similar attitudes toward embedded BMIs. More than half of the respondents, regardless of their disability status, had more negative attitudes toward the use of embedded BMIs than positive ones.

Furthermore, social concerns and ethical risks associated with the use of BMI devices, such as those related to privacy, human dignity, development of self-identity, and human relationships, were not as strongly identified by both groups as were operational and safety issues. Operational and security issues may be more readily identified or recalled, especially when the devices malfunction or are misused, social risks or ethical issues may be harder to conceive for those not using BMI devices in their daily lives. These observations corroborate several findings from our previous experimental qualitative surveys (see [4–7]) and may suggest that the long-term social risks and negative impacts of the use of cutting-edge cyborg technologies, such as BMIs, might still be difficult to imagine for individuals, regardless of their disability status. However, it is crucial to proactively identify the ethical and social risks associated with BMI use and consider countermeasures to mitigate them, because social impacts of them are irreversible.

On the other hand, the comparative analysis in this study highlighted some differences between respondents without disabilities and those with disabilities. Respondents without disabilities demonstrated motivation to use BMIs in various contexts and applications to envisage the use of BMI devices for preparing for anticipated emergencies, saving money and time, and assessing various forms of entertainment. However, these were not considered as critical requirements, and overall selection rates are also lower than for respondents with disabilities. In contrast, respondents with disabilities expressed a stronger desire to use cyborg technologies, including BMIs or BCIs, in more urgent, serious, and comprehensive ways to improve their lives and address their difficulties. In addition to their recognition of the beneficial aspects of BMIs, they also expressed greater awareness of risks and concerns than respondents without disabilities. In particular, the respondents with disabilities show a higher level of concern about the malfunction of BMIs and its physical impacts.

Interestingly, although no statistically significant differences could be found, one of the options regarding risk in the use of BMI, the concern of “Feeling as if I am being treated like an object”, was selected more frequently by the respondents without disabilities. This may imply that people without disabilities feel more anxious and uncomfortable about being cared for by machines or cyborg devices than those with disabilities. This insight could be explored as a topic for future discussion, as it should also be recognised as a concern for the ageing population.

5 Conclusions

In this study, questionnaire surveys were administered to two groups of people, one with disability and the other without disability, to investigate their expectations, possibilities, and risks of using BMIs and BCIs, and to compare the responses of the two groups. Considering the results and discussions of this study, it is suggested that there is a lack of opportunities to learn about and use cyborg devices, even for people with disabilities, whose need for cyborg technologies may be serious. Therefore, it is necessary to increase the awareness of cyborg technologies, including BMIs and the socio-technological environment in which these technologies are prevalent, among people with and without disabilities to further consider on the benefits and risks of these technologies. In addition, following questions should be discussed, such as how to proceed with the introduction of BMIs in an economically rational and socially appropriate manner, how to adapt the various benefits and risks associated with the use of BMIs to what purposes and availability, and how to address ethical issues and risks should be addressed in advance toward policy development on the use of BMIs.

Future research will include the analysis of responses that could not be introduced owing to volume limitations in this study; the analysis of other question items such as privacy, personal information, human relationship development, and artificial intelligence applications; quantitative analysis based on attributes such as whether differences by gender, age, type of disability, congenital or acquired disability and other attributes exist, and qualitative analysis of free descriptive responses in this survey. In addition, as the results of this survey were obtained from the Japanese respondents only, surveys in various countries and international comparative studies are required.

Acknowledgments. This study was supported by JSPS Grants-in-aid for Scientific Research 25K05673, 23K01545, and 22K02063, and by a research grant from the Telecommunications Advancement Foundation. We also appreciate the Center for Independent Living Hoshizora (Matsuyama, Ehime), all participants in the surveys, and the researchers who supported our study.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Shen, G., Horikawa, T., Majima, K., Kamitani, Y.: Deep image reconstruction from human brain activity. *PLoS Comput. Biol.* **15**(1), e1006633 (2019)
2. Chaudhary, U., Birbaumer, N., Curado, M.R.: Brain-machine interface (BMI) in Paralysis. *Ann. Phys. Rehabil. Med.* **58**(1), 9–13 (2015)
3. Liu, M., Ushiba, J.: Brain-machine Interface (BMI)-based neurorehabilitation for post-stroke upper limb paralysis. *Keio J. Med.* **71**(4), 82–92 (2022)
4. Orito, Y., et al.: The ethical aspects of a “psychokinesis machine: an experimental survey on the use of a brain-machine interface. In: Arias-Oliva, M., et al. (eds.) *Societal Challenges in the Smart Society Ethicomp* book series, pp. 81–91. Universidad de La Rioja, Logroño (2020)
5. Orito, Y., et al.: How a brain-machine interface can be helpful for people with disabilities? Views from social welfare professionals. In: Oliva, M.A., Borondo, J.P., Murata, K., Palma, A.M.L. (eds.) *Moving Technology Ethics at the Forefront of Society, Organisations and Governments*, pp.103–115 (2021)

6. Orito, Y., et al.: The social implications of brain machine interfaces for people with disabilities: experimental and semi-structured interview surveys. In: Proceedings of the ETHICOMP 2022: Effectiveness of ICT Ethics – How Do We Help Solve Ethical Problems in the Field of ICT? pp. 487–501 (2022)
7. Orito, Y., et al.: Is a brain machine interface useful for people with disabilities? Cases of spinal muscular atrophy. In: Proceedings of ETHICOMP 2024: The Leading Role of Smart Ethics in the Digital World, pp.9–19 (2024)
8. Cabinet Office: White paper on persons with disabilities (in Japanese). https://www8.cao.go.jp/shougai/whitepaper/r06hakusho/zenbun/siryo_01.html. Accessed 31 Dec 2024



Sustainable Fashion Trends on Second-Hand Shopping: The Role of Social Media

Lídia Monteiro¹, Orlando Lima Rua^{1,2} , Mar Souto-Romero³ ,
and Rafael Canorea-García³ 

¹ Porto School of Accounting and Business (ISCAP), Polytechnic Institute of Porto (P.PORTO),
Rua Jaime Lopes Amorim, s/n, 4465-004 S. Mamede de Infesta, Portugal

² Center for Organisational and Social Studies of the Polytechnic of Porto (CEOS.PP), Porto
School of Accounting and Business (ISCAP), 4465-004 S. Mamede de Infesta, Portugal

³ Department of Business Economics, Rey Juan Carlos University, 28032 Madrid, Spain
mar.souto@urjc.es

Abstract. This study aims to analyse the relation between sustainable fashion trends and second-hand shopping while keeping in mind the importance and role of social media. A qualitative methodology approach was adopted to analyse the research propositions and form a better understanding of the research gaps and further investigations. The main results highlight the influence of social media engagement, endorsement by influencers, transparent sustainability communication, and circular business models on second-hand shopping behaviours. The conclusion drawn underscore the significance of considering these factors in shaping consumer attitudes and behaviours towards sustainable fashion and second-hand shopping, while also acknowledging the limitations of qualitative research and offering suggestions for future research directions. Overall, this study contributes to a deeper understanding of the evolving landscape of sustainable fashion consumption and its implications for the fashion industry and society.

Keywords: Sustainable Fashion · Second-Hand Shopping · Social Media · Trends · University Students

1 Introduction

The growing awareness of environmental degradation and social injustices within the fashion industry has sparked a paradigm shift in consumer attitudes towards clothing consumption. Traditional notions of trend-driven fast fashion are increasingly being challenged by a burgeoning interest in sustainable alternatives, including second-hand shopping. This shift reflects a broader cultural movement towards conscious consumption and ethical living, where individuals are increasingly mindful of the environmental and social implications of their purchasing decisions [1].

Moreover, the pervasive influence of social media platforms has revolutionized the way in which fashion trends are disseminated and adopted [2]. With platforms such as

Instagram, TikTok, and Pinterest shaping consumer aspirations and preferences, understanding the role of social media in driving sustainable fashion practices becomes imperative. By leveraging the power of digital networks, individuals can not only access a vast array of second-hand offerings but also participate in online communities that promote sustainable lifestyles and values [3].

2 Literature Review

2.1 Sustainable Fashion Trends

Ahmad et al.'s [4] examines the knowledge, attitude, and practice (KAP) of adopting green fashion innovation, aiming to understand the factors influencing individuals in embracing sustainable practices within the fashion industry. A mixed-methods strategy is used in the research to collect both qualitative and quantitative data, and techniques through surveys, interviews, and focus group discussions. The key findings of the study are multifaceted. Firstly, it identifies various factors influencing the adoption of green fashion innovation, including environmental awareness, social influences, and personal values. The research highlights the varying levels of knowledge and awareness among participants, shedding light on gaps in understanding and areas for improvement in educational initiatives. Positive attitudes towards sustainable fashion were observed among participants, contributing to the likelihood of adopting green fashion practices. Additionally, the study identifies barriers to the widespread adoption of green fashion, such as the lack of accessibility to sustainable fashion products and economic constraints. In response to these findings, the researchers propose recommendations for improvement, including educational programs, increased accessibility to sustainable products, and awareness campaigns.

Johnstone and Lindh's [5] analyse the role of influencers in shaping sustainable fashion consumption behaviours among millennials in Europe. Employing a qualitative approach, the research relies on interviews and focus group discussions to glean insights from millennials regarding the influence of social media influencers on their fashion choices. Key findings underscore the substantial impact of influencers on millennials' fashion behaviours, emphasizing their role in promoting sustainable brands, encouraging unplanned buying behaviour, and serving an educational function. Trust and authenticity emerge as critical factors influencing the effectiveness of influencers, particularly in the context of sustainability.

De Oliveira, Miranda, and de Paula Dias [6] conduct a comprehensive exploration of customer perceptions of sustainable practices in both slow and fast fashion stores. The study combines quantitative surveys and qualitative interviews to capture a nuanced understanding of how customers perceive and evaluate sustainability initiatives in different fashion retail contexts. Key findings indicate varying levels of customer awareness regarding sustainability practices, with slow fashion stores generally perceived as more aligned with sustainability. Transparency in communication about sustainability initiatives emerges as crucial, influencing customer perceptions, although some express skepticism about the authenticity of such claims. The study reveals that sustainability practices significantly influence customers' purchasing decisions, particularly in slow

fashion stores, where customers are more likely to support brands aligned with their values. Educational initiatives, both in-store and online, are recognized as effective tools for enhancing customer awareness and understanding of sustainable practices. The findings suggest that customers value opportunities to learn about the environmental and ethical aspects of fashion.

Dragomir and Dumitru [7] study practical solutions for implementing circular business models in the fashion industry. With a focus on identifying strategies contributing to the success of circular approaches within the fashion supply chain, the study employs a practical and applied approach, utilizing case studies, industry examples, and best practices. The key findings emphasize the significance of closed-loop supply chains, where products undergo a continuous cycle of design, production, use, and recycling to minimize waste. Product design emerges as a crucial factor, advocating for designs that prioritize recyclability and reusability. Collaboration across the supply chain, involving manufacturers, retailers, and recycling facilities, is deemed essential for successful implementation. Engaging consumers in the circular economy through take-back programs and education initiatives is recognized, while technological innovations, including blockchain and data analytics, are highlighted for optimizing efficiency in circular supply chains. The practical implications underscore the importance of these findings for businesses seeking to adopt circular business models, offering actionable insights for sustainability and a roadmap for the fashion industry's transition towards circular practices.

2.2 Second-Hand Shopping

Valor et al.'s [8] contribute to the exploration of circular markets, focusing on second-hand clothing, by investigating the role of relational legitimacy in overcoming the stigma associated with such fashion choices. Employing a qualitative approach with case studies and in-depth interviews, the study aims to understand the factors influencing the acceptance and growth of circular markets. The study recognizes the existing stigma attached to second-hand clothing, emphasizing the imperative to address negative perceptions and stereotypes for the expansion of circular markets. It identifies relational legitimacy as a pivotal factor in overcoming this stigma. Building relationships and trust among stakeholders, including consumers, retailers, and policymakers, is crucial for legitimizing the second-hand clothing market. Moreover, the study emphasizes the role of supportive policies and regulations in legitimizing the circular market. Clear guidelines contribute to building trust and confidence in the second-hand clothing industry. Overall, the findings have implications for policymakers, retailers, and other stakeholders involved in the circular fashion market, providing insights into strategies for fostering the growth and acceptance of second-hand clothing markets.

Persson and Hinton [9] investigate the role of second-hand clothing markets in fostering a just circular economy. Employing a comprehensive approach, the study explores diverse business forms within the sector, encompassing traditional thrift stores, online platforms, and for-profit resale businesses. Profit motives emerge as significant influences, shaping the practices of second-hand clothing businesses and balancing profitability with ethical and sustainable considerations. The study examines into the contribution of second-hand clothing markets to a just circular economy, considering factors such

as equity, social justice, fair wages, working conditions, and impact on local communities. Exploring consumer perceptions and attitudes towards second-hand clothing, the research examines pricing, branding, and the overall shopping experience's influence on purchasing decisions.

Herold and Prokop [10] explore the potential of promoting rental clothing programs as a cost-effective and environmentally friendly substitute for rapid fashion. Utilizing a mixed-methods approach, the study explores motivations for engaging in rental fashion, including environmental concerns, a desire for variety, and a departure from fast fashion's disposable nature. Participants perceive sustainability in renting clothing, aligning with environmental values, and reducing the industry's ecological impact. Affordability emerges as a driving factor, allowing access to diverse wardrobes without the financial burden of fast fashion purchases. The research suggests a shift in consumer behaviour, emphasizing quality over quantity and valuing the experience of renting. The study discusses challenges and opportunities related to rental fashion schemes, including hygiene concerns and the potential to transform the industry toward sustainability.

2.3 Role of Social Media

Dwivedi et al.'s [11] sets out to explore and chart the future course of research in digital and social media marketing. With the collaborative efforts of multiple authors, it offers perspectives and research propositions to guide scholars and practitioners in navigating the evolving landscape of digital marketing. Employing a conceptual approach, the research synthesizes existing literature to identify emerging trends and propose future research directions. Key findings comprehend the identification of emerging trends in digital marketing, the integration of technologies such as artificial intelligence, and the examination of consumer behaviour in the digital age. Ethical considerations surrounding digital marketing, cross-cultural perspectives, and the influential role of digital influencers are also highlighted.

Shah and Asghar [12] investigation aim to research utilizing a social network representation to analyze the dynamics of social influence on consumer decisions. Employing a social network analysis approach, the study investigates the impact of network structure on consumption decisions, exploring factors contributing to the spread of preferences within social networks. Key findings include the influence of network topology, the role of homophily in shaping preferences, the significance of opinion leaders, and the temporal dynamics of consumption choices within networks. The search highlights the implications for marketing strategies, emphasizing the importance of understanding social networks for effective campaigns. It contributes to consumer behaviour literature by providing insights into the nuanced dynamics of consumption choices within the context of social networks, catering to both academia and practitioners.

Calderón-Fajardo et al. [13] aims to offer current insights into brand personality and delineate future research directions. The study explores the conceptualization, measurement, and impact of brand personality on consumer behaviour. Key findings include an examination of different conceptualizations of brand personality, methods for its measurement, its impact on consumer behaviour, and the influence of cultural and contextual

factors. The study also outlines future research directions, emphasizing the need for comprehensive measurement tools, exploration of brand personality in the digital age, and consideration of ethical dimensions in research.

2.4 Research Questions

Dwivedi et al. [11] highlight the evolving landscape of digital and social media marketing, a realm integral to conveying sustainable fashion trends and promoting second-hand shopping. The emergence of influencers play a pivotal role in disseminating sustainable fashion information and shaping consumer perceptions, potentially driving engagement with second-hand shopping [5].

Customer perceptions of sustainable practices in fashion stores contribute to understanding the factors influencing consumers in the context of sustainable fashion trends [6]. Transparent communication, a key aspect identified in the study, aligns with the goals of promoting second-hand shopping, where clarity about sustainability practices is essential in attracting consumers to pre-loved fashion items.

Second-hand clothing markets, including diverse business forms, profit motives, equity, and justice, provides critical insights into the societal and economic dimensions of sustainable fashion [9]. This aligns with the broader goals of sustainable fashion trends, emphasizing ethical considerations and social impact, which can contribute to the acceptance and promotion of second-hand shopping.

Social influence within social networks provides a lens through which to understand how sustainable fashion trends and second-hand shopping choices propagate through communities [9]. The role of opinion leaders and the dynamics of influence within social networks align with the influential nature of social media platforms, where trends are often set, and second-hand shopping can gain momentum.

Thus, emerging from the theoretical background, the student intends to test the following research questions (RQ):

RQ1 – Does social media engagement relates positively on Second-Hand Shopping?

RQ2 - Does consumer interest relies on the endorsement of Influencer's on Second-Hand Fashion?

RQ3 - Transparent Sustainability Communication brings a positive and significant effect on Second-Hand Fashion?

RQ4 – Composing a Circular Business Model creates a positive and significant effect on Second-Hand Fashion?

3 Methodology

Data collection was conducted through interviews. The interview script was created with the intention of gathering in depth perception on the influences of sustainable fashion trends on today's second-hand consumption and to better understand the role that social media plays in making decisions related to it. The selection of Gen Z consumers as the target demographic for this qualitative study is underpinned by their increasing influence in shaping consumer trends, particularly in the realm of sustainable fashion practices. As the first generation to come of age in a predominantly digital era, Gen Z individuals

are known for their strong digital literacy, social media engagement, and awareness of global issues, including environmental sustainability [14]. By including participants with similar educational backgrounds (students of the bachelor in Creativity and Innovation Business), the study aims to capture a range of perspectives and experiences and enhance the depth and comprehensiveness of the qualitative data collected. The sample has in total 5 participants that with different experiences created a nuanced standard that illustrated the results to the propositions.

The data was collected from April, 29th to May, 20th 2024, a thematic analysis approach was employed. Moreover, interpretive analysis was conducted to delve deeper into the underlying meanings and nuances inherent in the qualitative data. This qualitative analysis method facilitated a rich exploration of the research propositions, allowing for a nuanced understanding of the phenomena under investigation. Categories’ background can be seen on the table below (Table 1).

Table 1. Categories’ background.

Categories	Authors
Second-hand shopping habits and sustainable fashion awareness	Ahmad et al.’s (2020); De Oliveira, Miranda, and de Paula Dias (2022); Valor, Ronda, and Abril (2022); Johnstone and Lindh’s (2022); Shah and Asghar (2023)
Exploring the role of social media	Johnstone and Lindh’s (2022); Dwivedi et al.’s (2021); Shah and Asghar (2023)
Impact of sustainable fashion trends on second-hand shopping	De Oliveira, Miranda, and de Paula Dias (2022); Persson and Hinton (2023)

4 Results

The results provide valuable insights into the interconnected dynamics between sustainable fashion trends, second-hand shopping, and social media influence.

RQ1 – Does social media engagement relate positively on Second-Hand Shopping?

The findings from the interviews support RQ1, indicating a positive relationship between social media engagement and second-hand shopping. Participants consistently cited social media platforms, as influential channels for promoting sustainable fashion and encouraging second-hand shopping. The endorsements of influencers and the dissemination of content showcasing thrifted finds were perceived as effective in shaping consumer attitudes and behaviours towards second-hand shopping. Therefore, increased social media engagement correlates with greater interest and participation in second-hand shopping practices.

RQ2 – Does consumer interest rely on the endorsement of Influencers of Second-Hand Fashion?

RQ2 is also supported by the interview findings, highlighting the significant impact of influencer endorsements on consumer interest in second-hand fashion. Participants noted

that influencers play a crucial role in normalizing and popularizing second-hand shopping through their content on social media platforms. Recommendations from influencers were cited as influential factors driving consumer interest and engagement in second-hand shopping. Thus, the endorsement of influencers serves as a key motivator for consumers to explore and adopt second-hand fashion choices.

RQ3 – Transparent Sustainability Communication brings a positive and significant effect on Second-Hand Fashion?

While the interviewees did not directly address RQ3, the concept of transparent sustainability communication aligns with the overarching theme of environmental awareness and ethical consumption discussed by participants. Transparent communication regarding sustainability practices, such as promoting the environmental benefits of second-hand shopping, could potentially have a positive effect on consumer perceptions and behaviours. However, further empirical research focusing specifically on the impact of transparent sustainability communication on second-hand fashion adoption is warranted to validate this proposition.

RQ4 – Composing a Circular Business Model creates a positive and significant effect on Second-Hand Fashion?

RQ4 suggests that composing a circular business model positively influences second-hand fashion. Although not explicitly explored by the interviewees, the concept of a circular economy underpins the principles of sustainability and resource efficiency, which are inherent to second-hand fashion practices. Therefore, it can be inferred that the adoption of circular business models, such as upcycling, recycling, and resale, contributes to the promotion and growth of second-hand fashion. Future research could delve deeper into the direct impact of circular business models on second-hand fashion adoption to validate this proposition empirically.

The findings from the interviews provide valuable insights into the dynamics between social media engagement, influencer endorsements, sustainability communication, and circular business models in relation to second-hand fashion. These propositions offer avenues for further research and practical implications for stakeholders seeking to promote and support sustainable fashion practices, particularly within the context of second-hand shopping.

5 Conclusion

This study provides a nuanced exploration of the interplay between sustainable fashion trends, social media, and second-hand shopping behaviors. Through qualitative interviews, participants shared valuable insights into their perceptions, motivations, and behaviours related to sustainable fashion consumption and second-hand shopping practices. The findings underscore the multifaceted nature of these dynamics, highlighting the pivotal role of social media platforms, influencers, transparent communication strategies, and circular business models in shaping consumer attitudes and behaviors. By illuminating these complexities, this research contributes to a deeper understanding of the evolving landscape of sustainable fashion consumption and its implications for the fashion industry and society at large.

Acknowledgments. This research was funded by the Telefónica Chair on Smart Cities at Universitat Rovira i Virgili and Universitat de Barcelona (project number 42.DB.00.18.00).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Joy, A., Sherry, J.F., Jr., Venkatesh, A., Wang, J., Chan, R.: Fast fashion, sustainability, and the ethical appeal of luxury brands. *Fash. Theory* **16**(3), 273–296 (2012). <https://doi.org/10.2752/175174112X13340749707123>
2. Kim, A.J., Ko, E.: Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. *J. Bus. Res.* **65**(10), 1480–1486 (2012). <https://doi.org/10.1016/j.jbusres.2011.10.014>
3. Caro, F., Martínez-de-Albéniz, V.: Fast fashion: business model overview and research opportunities. In: Agrawal, N., Smith, S.A. (eds.) *Retail Supply Chain Management: Quantitative Models and Empirical Studies*, vol. 223, pp. 237–264. Springer, Boston (2015). https://doi.org/10.1007/978-1-4899-7562-1_9
4. Ahmad, A., Madi, Y., Abuhashesh, M., Nusairat, N.M., Masa'deh, R.: The knowledge, attitude, and practice of the adoption of green fashion innovation. *J. Open Innov. Technol. Market Complex.* **6**(4), 107 (2020). <https://doi.org/10.3390/joitmc6040107>
5. Johnstone, L., Lindh, C.: Sustainably sustaining (online) fashion consumption: using influencers to promote sustainable (un)planned behaviour in Europe's millennials. *J. Retail. Consum. Serv.* **64**, 102775 (2022). <https://doi.org/10.1016/j.jretconser.2021.102775>
6. De Oliveira, L.G., Miranda, F.G., de Paula Dias, M.A.: Sustainable practices in slow and fast fashion stores: What does the customer perceive? *Clean. Eng. Technol.* **6**, 100413 (2022). <https://doi.org/10.1016/j.clet.2022.100413>
7. Dragomir, V.D., Dumitru, M.: Practical solutions for circular business models in the fashion industry. *Clean. Logist. Supply Chain* **4**, 100040 (2022). <https://doi.org/10.1016/j.clscn.2022.100040>
8. Valor, C., Ronda, L., Abril, C.: Understanding the expansion of circular markets: building relational legitimacy to overcome the stigma of second-hand clothing. *Sustain. Prod. Consum.* **30**, 77–88 (2022). <https://doi.org/10.1016/j.spc.2021.11.027>
9. Persson, O., Hinton, J.B.: Second-hand clothing markets and a just circular economy? Exploring the role of business forms and profit. *J. Clean. Prod.* **390**, 136139 (2023). <https://doi.org/10.1016/j.jclepro.2023.136139>
10. Herold, P.I., Prokop, D.: Is fast fashion finally out of season? Rental clothing schemes as a sustainable and affordable alternative to fast fashion. *Geoforum* **146**, 103873 (2023). <https://doi.org/10.1016/j.geoforum.2023.103873>
11. Dwivedi, Y.K., et al.: Setting the future of digital and social media marketing research: perspectives and research propositions. *Int. J. Inf. Manage.* **59**, 102168 (2021). <https://doi.org/10.1016/j.ijinfomgt.2020.102168>
12. Shah, S.S., Asghar, Z.: Dynamics of social influence on consumption choices: a social network representation. *Heliyon* **9**(6), e17146 (2023). <https://doi.org/10.1016/j.heliyon.2023.e17146>
13. Calderón-Fajardo, V., Molinillo, S., Anaya-Sánchez, R., Ekinci, Y.: Brand personality: current insights and future research directions. *J. Bus. Res.* **166**, 114062 (2023). <https://doi.org/10.1016/j.jbusres.2023.114062>
14. Palomo-Domínguez, I., Elías-Zambrano, R., Álvarez-Rodríguez, V.: Gen Z's motivations towards sustainable fashion and eco-friendly brand attributes: the case of vinted. *Sustainability* **15**(11), 8753 (2023). <https://doi.org/10.3390/su15118753>



The Nexus Between Artificial Chatbots and Individual Creativity

Luana Coelho¹, Orlando Lima Rua^{1,2,3} , and Felipe Debasa⁴  

¹ Porto School of Accounting and Business (ISCAP), Polytechnic Institute of Porto (P.PORTO), Rua Jaime Lopes Amorim, S/N, 4465-004 S. Mamede de Infesta, Portugal

² Center for Organisational and Social Studies of the Polytechnic of Porto (CEOS.PP), Porto School of Accounting and Business (ISCAP), 4465-004 S. Mamede de Infesta, Portugal

³ Research Center of Business Sciences (NECE), University of Beira Interior (UBI), 6200-209 Covilhã, Portugal

⁴ Department of Business Economics, Rey Juan Carlos University, 28032 Madrid, Spain
felipe.debasa@urjc.es

Abstract. This study aims to analyse the relationship between the use of artificial intelligence Chatbots and individual creativity. A quantitative methodological approach was adopted using an online survey with 186 students from the higher education institution of the Porto School of Accounting and Business (ISCAP), of the Polytechnic Porto (Portugal). The results indicate that students who considered themselves to be creative don't show intention of using AI chatbot services, despite finding them useful. This study seeks to encourage further academic and scientific exploration in a fast-growing and emerging field that is increasingly gaining relevancy in our society. As artificial intelligence tools become more common in our daily lives, the results of such research will continue to change. Consequently, future studies in this area are likely to produce different findings, highlighting the ever-evolving nature of this emerging field. This study affirms that the impact of artificial intelligence on the contemporary world is already substantial, driving significant transformations across multiple sectors, without the need for speculation about its future developments.

Keywords: Artificial Intelligence · Chatbots · Individual Creativity · SEM

1 Introduction

An artificial language model is a powerful computational tool that leverages the latest advances in artificial intelligence and machine learning algorithms to process, understand, and generate natural-sounding human language text that is both coherent and grammatically correct [1]. Within the domain of AI, machine learning is a subfield that is dedicated to developing software that can learn and evolve independently, ultimately leading to improved performance over time. [2] sustain that these language models utilize sophisticated algorithms that can identify patterns in the vast amounts of data inputs they receive, enabling the models to make informed decisions and predictions. AI language

models are typically built upon machine learning techniques, including deep learning neural networks, which have been shown to significantly improve their accuracy and efficiency overtime [3]. The success of modern AI language models is largely attributed to the powerful capabilities of machine learning, which enables these models to continually refine their understanding of complex problems and adapt to new situations with increasing accuracy [4]. These models are used in a variety of applications, such as chatbots, virtual assistants, language translation, and text generation.

This segmentation of university students carries significant structural implications for the relationship between individuals and technology within academic contexts. On one hand, it facilitates the anticipation of asymmetric patterns of technological appropriation, which may result in varying levels of functional dependence on automated systems.

In this context, the relationship between university students and technology—particularly artificial intelligence systems such as chatbots—should be examined from the standpoint of the present moment. Understanding this interaction is crucial for analyzing the disruptive transformations these technologies are already prompting in cognitive practices, educational processes, and, by extension, the structural configuration of contemporary society within the framework of the current world.

2 Literature Review

A chatbot is a sophisticated computer program designed to simulate human conversation and interact intelligently with users [5]. This remarkable ability to mimic human-like conversation is made possible through a powerful technology known as Natural Language Processing (NLP) [6]. NLP is a subfield of Artificial Intelligence intertwined with linguistics that enables computers to comprehend and interpret written or spoken human languages [7, 8]. Current technologies empowers chatbots to understand the nuances of statements or words expressed in natural language, bridging the gap between humans and machines in communication and fostering a more intuitive and interactive user experience [9].

Individual creativity is a fundamental human capacity that enables people to generate novel and valuable ideas, solutions, and products [10]. According to [11], creativity is identified as an individual characteristic and is primarily studied in the field of psychology. This ability can manifest itself in all kinds of ways and is used for a multitude of purposes since it is rooted in the cognitive process of an individual, and it plays an important role in accomplishing small everyday tasks to significant achievements [12].

In recent years, the interaction between humans and chatbots has evolved rapidly, both in form and function. Early chatbot systems were limited to executing predefined commands through linear, text-based responses, lacking adaptive or contextual understanding. Today, however, most conversational agents are built upon sophisticated Natural Language Processing models and increasingly incorporate voice interfaces designed to emulate human communication with high degrees of fluency and coherence. These advances enable the simulation of structured dialogue, approaching levels of complexity that challenge traditional distinctions between human and machine interaction.

This technological shift toward a critical disruption point that aligns with the concept of the Technological Singularity, originally proposed by Alan Turing. The Singularity refers to a hypothetical moment in which artificial intelligence systems would surpass human cognitive capacities. A precondition for this scenario, as Turing suggested, involves the successful passing of the Turing Test, in which a human interlocutor is unable to determine whether they are conversing with a human or a machine. While this remains a theoretical horizon, the current capabilities of artificial intelligence are already producing significant transformations in social behavior, communication, and cognitive dynamics. Within a transhumanist framework, these technologies are understood as functional extensions of human faculties, both physical and mental. Such a perspective requires rethinking the ethical, legal, and educational frameworks that govern the integration of intelligent systems into everyday life. The social and cognitive implications of AI-driven systems are increasingly being addressed across multiple disciplines. However, when examined through the lens of interdisciplinary and historically informed inquiry, they also intersect with the emerging domain of the History of the Present.

3 Methodology

Data was collected from a students' online survey of the Porto School of Accounting and Business (ISCAP), of the Polytechnic Porto (Portugal). The survey link was distributed to the students through various channels, including social media platforms (Facebook, Twitter and Instagram), and e-mail. By leveraging the reach and connectivity of these channels, the survey reached a variety of students in a timely manner. The data collection period was between 14th and 29th of May, 2023.

To assess the use of AI chatbots, pre-established scales were employed. Social influence (SI) was measured using three items [13]. Chatbot usage intention and performance expectancy (CB) were examined using three items each [14]. Compatibility (C) was measured with the adoption of six items [15]. The previous scales were adapted to meet the population, since the questionnaire is targeted to students. To measure individual creativity (IC), we drew upon the adaptation of the creativity scale initially developed by [16], with subsequent modifications introduced by [17], and [18]. The scale comprises five items designed to gauge respondents' proficiency in generating novel ideas, embracing innovation, experimenting with new concepts, offering innovative solutions, and serving as a role model for creativity. All items were presented on seven-point Likert-type scales (1 - strongly disagree to 7 - strongly agree).

4 Results

PLS-SEM was used to test the hypotheses with SmartPLS 3.0 software [19]. PLS-SEM was best suited to estimate the research model as: (1) this research focuses on prediction and explanation of the variance of the model's constructs (in this case, four); (2) the research model has a complex structure; (3) the relationship between open innovation and competitive advantage can be measured directly and indirectly; (4) the study uses first and second-order reflective constructs; and (5) the sample ($n = 251$) is relatively small.

The results showed that the measurement model met all general requirements (Table 1).

Table 1. Standardised factor analysis loadings, CR, AVE, and HTMT.

First-order constructs	Cronbach's Alpha (> .7)	CR (> .7)	AVE (> .5)	HTMT			
				C	CB	IC	SI
Compatibility (C)	.822	.880	.650				
Chatbot usage intention and performance expectancy (CB)	.783	.863	.678	.353			
Individual creativity (IC)	.937	.955	.841	.396	.454		
Social influence (SI)	.704	.780	.543	.699	.846	.772	

SI had a significant and positive relationship with IC ($\beta = 0.111$, $t = 1.561$; $LL = 0.437$, $UL = 0.621$); thus, H1 was supported. The CB had a significant and positive effect on IC ($\beta = 0.121$, $t = 1.589$); thus, H2 was supported. Moreover, C had a significant and positive effect on IC ($\beta = 0.492$, $t = 6.119$); thus, H3 also was supported. Table 2 and Fig. 1 shows the mentioned results.

Table 2. Results of the structural equations modeling.

Hypotheses	Original Sample (O)	Sample Mean (M)	Standard Error (STERR)	T-Statistics (O / STERR)	Results
H1. SI-- > IC	.111	.126	.071	1.561**	Supported
H2. CB-- > IC	.121	.123	.076	1.589**	Supported
H3. C -- > IC	.492	.489	.080	6.119*	Supported

Notes: * $p < 0.001$; ** $p < 0.05$

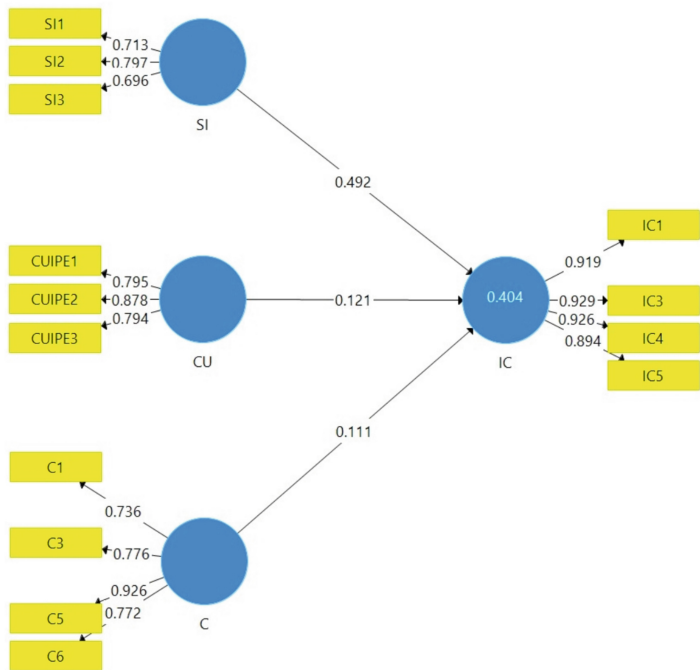


Fig. 1. Structural model assessment.

5 Conclusion

Upon analyzing the findings of the study, the majority of the population expresses disagreement regarding their intention to use AI chatbot services. This observation aligns with the previous paragraph since the larger portion of the respondents positions themselves favorably on the creative scale, indicating that individuals who exhibit higher levels of creativity may rely less on the assistance of AI chatbots during the initial stages of their creative processes. These findings highlight the intricate relationship between individual creativity and the perceived need for AI chatbot support. They indicate that highly creative individuals are more likely to rely on their own cognitive abilities and ideation strategies, potentially diminishing the perceived necessity of AI chatbots in their creative journey. Furthermore, the prevailing disagreement among the sampled population underscores the disparity between the perceived benefits of AI chatbots and individual's personal preferences.

The responses regarding social influence and the use of AI chatbots were characterized by a significant division between neutrality and disagreement. These findings indicate that the topic of AI chatbot usage lacks relevance within social and familial circles. This observation holds considerable weight when considering individuals' intention to use AI chatbots. If we are not motivated or encouraged by our peers and social circles on a certain matter, it becomes increasingly unlikely that we will actively engage with it. However, a discrepancy arises when comparing the responses regarding the perceived usefulness of chatbots with the intentions to use them. Despite a majority

of participants expressing agreement with the notion of chatbot usefulness, the section of the questionnaire regarding the use of AI chatbots reveals that the vast majority expresses uncertainty or disagreement about their intention to employ chatbot services.

The findings of the study show a notable division of opinions among participants regarding the compatibility of chatbot services with their preferred methods of performing creative and intellectual tasks. The survey responses revealed a range of sentiments, spanning from uncertainty to complete disagreement. This divergence in viewpoints indicates that the sampled population has reservations and hesitations when considering the integration of AI chatbots into their creative and intellectual processes. These contrasting perspectives warrant further exploration to uncover the underlying reasons behind these discrepancies.

The findings of the study show a notable division of opinions among participants regarding the compatibility of chatbot services with their preferred methods of performing creative and intellectual tasks. The survey responses revealed a range of sentiments, spanning from uncertainty to complete disagreement. This divergence in viewpoints indicates that the sampled population has reservations and hesitations when considering the integration of AI chatbots into their creative and intellectual processes. These contrasting perspectives warrant further exploration to uncover the underlying reasons behind these discrepancies.

The study reveals a significant dissonance between the perceived usefulness of chatbots and their actual adoption, particularly between students who identify as creative and those who do not. Based on the results obtained, the need to incorporate critical digital literacy in artificial intelligence within university educational environments becomes evident. Such training would undoubtedly help prevent a new form of digital divide—this time concerning AI applications. The resistance of more creative students to using chatbots, despite acknowledging their usefulness, points to a lack of cognitive and pedagogical tools necessary to assess the real role of these technologies in intellectual processes, particularly in the context of educational innovation. Higher education institutions must address, in a transversal manner, the development of reflective competencies, reading comprehension, and critical thinking. These reflective abilities are essential to evaluate the role of artificial intelligence in contrast to original thinking, complex reasoning, and symbolic production. This divergence may lead to differentiated educational trajectories in terms of creativity development, problem-solving skills, and the consolidation of authorship. Far from suggesting an incompatibility between creativity and artificial intelligence, the findings call for a reconsideration of the necessary balance between technological augmentation and cognitive sovereignty.

The absence of significant social influence in chatbot usage, as indicated by responses related to peer pressure or validation, emphasizes that technological adoption cannot be explained solely by individual factors. Instead, it is shaped by cultural, social, and relational dimensions. Technological decisions are embedded within a framework of social norms, shared identities, and generational narratives that require deeper investigation. From the perspective of Contemporary History, this research documents a transitional moment in the relationship between human cognition and artificial systems. The emergence of intelligent assistants capable of simulating creative behavior presents not only technical challenges but also ethical questions.

Several avenues for future research emerge from this study. It would be pertinent to develop studies that examine how attitudes toward chatbots evolve over time and in different contexts. Comparative analyses across academic disciplines, generations, or sociocultural environments would also be of great interest. Finally, qualitative approaches could enrich these findings through in-depth analysis of the beliefs, emotions, and interpretive frameworks underlying the acceptance or rejection of these emerging technologies.

Acknowledgments. This research was funded by the Telefónica Chair on Smart Cities at Universitat Rovira i Virgili and Universitat de Barcelona (project number 42.DB.00.18.00).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Goldberg, Y.: A primer on neural network models for natural language processing. *J. Artif. Intell. Res.* **57**, 345–420 (2019)
2. Dumitriu, D., Popescu, M.A.: Artificial intelligence solutions for digital marketing. *Procedia Manuf.* **46**, 630–636 (2020)
3. Jurafsky, D., Martin, J.H.: Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Pearson Education (2019)
4. Manning, C.D.: Computational linguistics and deep learning. *Ann. Rev. Linguist.* **6**, 365–387 (2020)
5. Barnes, S.J., Vidgen, R.T.: The impact of chatbots on customer service: a systematic review. *J. Serv. Manag.* **29**(2), 233–256 (2018)
6. Khanna, A., Pandey, B., Vashishta, K., Kalia, K., Bhale, P., Das, T.: A study of today's A.I. through chatbots and rediscovery of machine intelligence. *Int. J. u- and e-Serv. Sci. Technol.* **8**, 277–284 (2015)
7. Adamopoulou, E., Moussiades, L.: An overview of chatbot technology. In: *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, 5–7 June 2020, Proceedings, Part II*, vol. 584, pp. 373–383 (2020)
8. Ramesh, V., Singh, A.: Chatbots: a comprehensive overview. *J. Big Data* **4**(1), 1–22 (2017)
9. Hughes, D.J., Lee, A., Tian, A.W., Newman, A., Legood, A.: Leadership, creativity, and innovation: a critical review and practical recommendations. *Leadersh. Q.* **29**(5), 549–569 (2018)
10. Torrance, E.P.: Why teach thinking? Creativity and giftedness. *Gifted Child Today* **25**(3), 48–57 (2002)
11. Mumford, M.D., Gustafson, S.B.: Creativity syndrome: integration, application, and innovation. *Psychol. Bull.* **103**(1), 27–43 (1988)
12. Oliveira, T., Faria, M., Thomas, M.A., Popovic, A.: Extending the understanding of mobile banking adoption: when UTAUT meets TTF and ITM. *Int. J. Inf. Manag.* **34**(5), 689–703 (2014)
13. Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User acceptance of information technology: toward a unified view. *MIS Q.* **27**(3), 425–478 (2003)

14. Moore, G.C., Benbasat, I.: Development of an instrument to measure the perceptions of adopting an information technology innovation. *Inf. Syst. Res.* **2**(3), 192–222 (1991)
15. Tierney, P., Farmer, S.M., Graen, G.B.: An examination of leadership and employee creativity: the relevance of traits and relationships. *Pers. Psychol.* **52**, 591–620 (1999)
16. Farmer, S.M., Tierney, P., Kung-Mcintyre, K.: Employee creativity in Taiwan: an application of role identity theory. *Acad. Manag. J.* **46**(5), 618–630 (2003)
17. Lin, C., Wittmer, J.L.S., Luo, X.: Cultivating proactive information security behavior and individual creativity: the role of human relations culture and IT use governance. *Inf. Manag.* **59**(6), 103650 (2022)
18. Hair, J., Hult, G.T.M., Ringle, C., Sarstedt, M.: *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Sage, London (2016)



Symbolic Aspects of Online Privacy Protection Behaviour: From a Social Communication Perspective

Yasunori Fukuta¹ , Kiyoshi Murata¹ , Yohko Orito² ,
Vanessa Bracamonte³ , and Takamasa Isohara³ 

¹ Meiji University, Chiyoda 101-8301, Tokyo, Japan
yasufkt@meiji.ac.jp

² Ehime University, Matsuyama 790-8577, Ehime, Japan

³ KDDI Research, Inc., Fujimino 356-8502, Saitama, Japan

Abstract. This study provides a conceptual review of the symbolic dimensions of privacy protection behaviours (PPBs) in social media communication, drawing on interactionist frameworks, particularly those of Blumer and Goffman. It also investigates their core characteristics through evidence gathered from semi-structured interviews and a questionnaire survey. Our findings show that the symbolic framing of explicit PPBs is linked to the emergence of potentially undesirable self-impressions, such as appearing overly preoccupied with privacy. This, in turn, may discourage individuals from adopting such behaviours. These findings contribute to a better understanding of the symbolic barriers to privacy-related behaviours in mediated social contexts.

Keywords: Privacy Protection Behaviour · Symbolic Aspects of Privacy Protection Behaviour · Social Communication · Privacy Paradox · Social Media Users

1 Introduction

In the digital age, privacy risks are pervasive, taking multiple forms within the highly developed ecosystem of personal data. Accordingly, achieving a comprehensive understanding and effective mitigation of these risks remains a significant challenge [1]. Many national legal systems are based on notice-and-consent regimes, embedding the principle of “privacy self-management” into a range of privacy practices [2, 3]. Within this framework, organisations handling personal data are obliged to provide adequate prior notice of their data practices, while individuals are expected to manage their own privacy risks on the basis of the information provided.

Privacy is thus discussed both as a right to be protected through legal or institutional mechanisms and as a personal responsibility to be safeguarded by individuals. In the context of social media communication in particular, privacy is framed as something each user is expected to maintain voluntarily. Numerous studies have already identified

concrete protective measures, such as excluding contact information, limiting profile visibility, untagging or deleting photos, using anti-spyware tools, and employing anonymity or encryption software, and have examined how these are used to address social privacy threats [4, 5].

Privacy protection behaviour (PPB) has typically been conceptualised as a rational strategy for mitigating risk. A widely used theoretical framework for explaining PPB is Protection Motivation Theory [6, 7]. Originally proposed by Rogers [8] and subsequently extended, this theory views PPB decision-making as a function of threat appraisal (e.g., severity and likelihood) and coping appraisal (e.g., perceived efficacy of the behaviour and self-efficacy). In this view, PPB is framed as a rational response to perceived risks and the available means of mitigation.

Empirical studies confirm that social media users often exhibit rational PPBs [5]. However, research on the privacy paradox suggests that actual behaviour frequently deviates systematically from what would be considered rational [9]. This is especially evident on social media, where PPBs occur within specific interactional contexts, such as sharing updates, interacting in limited social networks, or conversing with close friends. Given this, it is plausible that PPBs may function not only as rational responses but also as communicative acts, conveying implicit messages to intended audiences. Drawing on the interactionist notion that human behaviour constitutes a form of social communication [10, 11], this study explores the symbolic dimensions of PPBs through semi-structured interviews and a large-scale questionnaire survey. These symbolic aspects may help to explain the systematic deviations from rationality observed in PPBs on social media, offering valuable insights into the privacy paradox. They may also hold important implications for the design of legal frameworks, technologies, and services related to privacy protection.

2 The Symbolic Aspects of PPB Among Social Media Users

2.1 Theoretical Background

To examine the symbolic dimensions of PPB, this study draws on Herbert Blumer's theory of symbolic interactionism and Erving Goffman's dramaturgical approach. Blumer conceptualised individuals as agents who derive meanings for things—including objects and actions—through social interaction, and who act based on those meanings [10]. In this framework, behaviour is not guided by an a priori system of meanings, but emerges from interactions with others, shaped by meanings co-constructed in the process. Here, 'interaction' goes beyond repeated sequences of actions and reactions between independent agents. In symbolic interactionism, it entails mutual symbolic constraint, whereby meaning is co-constructed. More specifically, interaction between individuals *A* and *B* occurs when each sees the other as someone who acts with reciprocal awareness—*A* considers *B*'s perspective, and *B* does the same for *A*. Any self-attributed meaning thus anticipates the other's interpretation and response.

Goffman's dramaturgy provides a complementary analytical lens [11, 12]. It interprets everyday social interaction through theatrical metaphors, where individuals are seen as performers presenting themselves on a "front stage" to create desired impressions on their audience, i.e., their interlocutors. In such performances, actions are oriented

towards the audience, guided less by immediate feedback than by assumptions about how one's behaviour will be perceived. Impression management, then, is anticipatory in nature.

Building on these ideas, this study conceptualises PPB not solely as a functional privacy measure but as an act carrying symbolic meaning, where users are attuned to the impressions their actions create. This impression-oriented perspective constitutes what is referred to here as the symbolic aspect of PPB.

However, the theoretical roots of these concepts lie primarily in the mid-20th century, and much of this work assumes face-to-face interaction as its context. Hence, adaptations—or even significant revisions—may be needed when applying these theories to social media. Face-to-face interaction involves direct interpersonal orientation and synchronous engagement [13]. For instance, a conversation in a café unfolds in real time between co-present individuals. Both performer and audience share the same physical and temporal setting, making the interaction highly intentional and directed. Goffman terms this “focused interaction” [14].

By contrast, Bernie Hogan suggests that social media interaction is better understood through the notion of exhibition [13]. Social media posts are not fleeting; they persist, allowing asynchronous viewing and response. Unlike email, they often lack a strong interpersonal focus. Even when shared with a restricted audience, content is broadcast in a consistent tone to all recipients. Hogan thus argues that “exhibition” better captures the nature of social media interactions than “performance”.

Further insight is offered by David R. Brake's analysis of computer-mediated communication [12]. He highlights the paradox of online communication: it allows potential mass reach while also facilitating direct replies. This combination contributes to a lack of audience transparency. Users must engage in impression management without knowing who their audience truly is. Due to missing contextual cues, they often rely on generalised assumptions and must anticipate reactions from an audience that remains ambiguous and abstract.

From these perspectives, the symbolic aspect of PPB on social media can be understood as a form of impression management that diverges significantly from its real-time, interpersonal, and context-rich counterpart in face-to-face communication.

2.2 Explicit PPB in Controlled Communication Spaces

Interactions on social media encompass a wide range of content and contexts. From the perspective of privacy protection, these interactions are often categorised according to whether they occur via a public or private user account. In this framework, the former is commonly associated with Goffman's concept of the front stage, while the latter is positioned as the back stage. However, communication through private accounts also demonstrably exhibits front-stage characteristics [13]. Even when users interact with audiences they have personally approved, such exchanges can still entail privacy risks. Moreover, these audiences may span a diverse array of relational contexts (e.g., professional contacts, hobby-based acquaintances, or childhood friends) and levels of intimacy (ranging from close personal relationships to more distant ties). Precisely because of this relational variability, users are often motivated to protect particular aspects of their privacy and manage the impressions they convey to others.

This study focuses on Japanese university students. In Japanese culture, self-presentation is closely shaped by how individuals subjectively position others in relation to themselves. Adams, Murata, and Orito [15] note that the boundary between *uchi* (inside—denoting close relations) and *soto* (outside—referring to more distant others) is fluid, with a liminal zone between the two. Since followers on private social media accounts are typically approved connections, they rarely fall within the *soto* category. Nevertheless, as they often span varying degrees of familiarity, interactions within private accounts may occur not only within the *uchi* sphere but also in this liminal space. This context aligns with the impression management challenges posed by the “ambiguous and abstract” audiences characteristic of social media, as discussed in Sect. 2.1.

Private accounts, therefore, provide a particularly appropriate setting for examining the symbolic dimension of PPB. On this basis, the present study centres on information disclosure and PPB in the context of private social media use.

A range of PPBs and related strategies in social media settings has been the subject of ongoing discussion. These include excluding contact information, limiting profile visibility, untagging or deleting photos, adjusting privacy settings, and deploying technical tools such as anti-spyware software, anonymity services, or encryption—though some of these measures may overlap in function or intent [4, 5]. Backstage PPBs such as adjusting privacy settings or installing anti-spyware software are generally invisible to communication partners on the front stage, and thus carry limited symbolic significance. By contrast, strategies that make privacy protection clearly observable to others may be understood as symbolic. This study focuses on an explicitly visible PPB: the posting or sharing of photos with edited faces or backgrounds.

In the following sections, building on the preceding theoretical discussion and analytic focus, we empirically examine the symbolic dimensions of PPBs among social media users, drawing on findings from semi-structured interviews and a large-scale questionnaire survey.

3 Exploratory Insights from the Semi-structured Interviews

Building on the preceding conceptual discussion, this study conducted an exploratory analysis of how social media users perceive their own acts of explicit PPB when using private accounts. The analysis was based on semi-structured interviews carried out via Zoom between October and December 2024. Both this interview study and the questionnaire survey discussed in the following section received ethical approval from the committees of Meiji University and Ehime University.

Participants were undergraduate students at Meiji University: nine individuals aged 19 to 22 (six male, three female). The initial two interviews revealed that the features and functions of social media platforms vary considerably across services, shaping communicative behaviour. Consequently, from the third interview onwards, the study focused on Instagram—a platform widely used for photo sharing—and limited recruitment to participants with experience posting photos via a private Instagram account. Participants were first asked about their SNS usage and its purposes, followed by questions on their perceptions and feelings when posting or encountering photos that had been edited to obscure facial features or backgrounds.

All nine participants reported that they did not use open Instagram accounts and maintained multiple private ones. Posting frequency varied; even frequent users often used certain accounts primarily to view others' content. This suggests a general tendency to differentiate usage by account. Regarding photo sharing, most preferred Instagram Stories—which disappear after a short period—over standard feed posts.

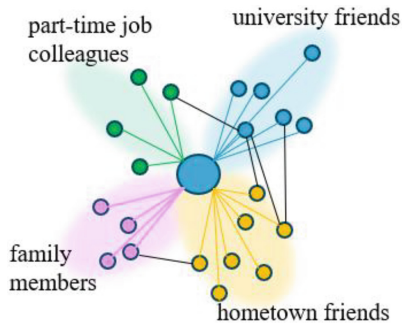


Fig. 1. The network within private accounts as an ego-centred, multi-relationship network

It also emerged that participants formed layered interactions within their private accounts, distinguishing between posts visible to all approved followers and those shared only with “Close Friends”. Moreover, regardless of relational closeness, single private accounts often included diverse social ties (e.g., hometown friends, family, university peers, part-time colleagues). This network can be characterised as an ego-centred, multi-relational structure (see Fig. 1).

Three key themes emerged regarding perceptions and feelings about posting or encountering edited photos. First, users appear sensitive to the symbolic aspects of explicit PPB. For instance, participant *A* (male, 22) noted that although he used to post edited photos (e.g., with stickers or mosaic effects on faces) in high school, he no longer does so: “Since people around me stopped editing photos, it now feels awkward to do so. ... Everyone else stopped, so I felt I shouldn’t either.” This suggests that editing practices are influenced by perceived social norms and others’ evaluations.

Participant *B* (male, 21), who primarily used Instagram for limited friend interactions, remarked: “I can’t help but wonder why they edited it. It feels like they’re overly concerned about privacy, and I don’t want to be seen that way by others.” Despite his relaxed attitude toward privacy risks, he interpreted photo edits as conveying meanings, such as “excessive concern.”

Such responses suggest that users are attuned to how visible PPBs are interpreted by others and how they align with social norms. This points to a symbolic dimension of explicit PPBs, with implications for self-presentation and impression management.

Second, these symbolic aspects appear especially salient in limited communication spaces such as private Instagram accounts. Participant *C* (female, 21) remarked, “If I received a photo with edits on my private account, I’d feel like they don’t trust me. ... I wouldn’t like it. I share my photos because I trust them.” Similarly, participant *D* (male, 21) stressed mutual trust: “When choosing a photo to send, I consider whether it’s one

I wouldn't need to edit. I think the same goes for others." These views suggest that the symbolic dimension becomes more pronounced not in public interactions, but in private ones grounded in interpersonal relationships.

Finally, as noted in several responses, symbolic meanings linked to explicit PPBs often carry negative connotations, such as appearing "overly concerned with privacy" or implying "a lack of trust." From an impression management perspective, individuals may avoid such behaviours to prevent being negatively perceived. Thus, the symbolic dimension may consciously or unconsciously discourage users from adopting explicit PPBs, potentially undermining privacy protection.

4 Results of the Quantitative Survey

4.1 Overview of the Survey

To assess the broader relevance and generalisability of the findings derived from the semi-structured interviews, a questionnaire survey was conducted. Administered between 12 and 20 April 2025, the survey targeted university students enrolled at Meiji University, Ehime University, and Toyama University in Japan.

There were two principal reasons for focusing exclusively on university students. First, they represent the digital native generation, who regularly engage in social media-based communication, thereby ensuring continuity with previous research that has focused on similar populations. Second, restricting the sample to this group allowed for greater control over variability stemming from individual differences. Given that communication practices are likely shaped by generational and socio-positional factors, it was considered that the advantage of reducing such variability outweighed the drawback of limiting generalisability. Nonetheless, the authors fully acknowledge that evaluating the applicability of these findings to wider demographic groups remains a pertinent topic for future investigation.

Respondents accessed the survey via Google Forms, where they were provided with an overview and asked to give informed consent. Of the 748 individuals who accessed the form, 701 consented and completed questions concerning their social media usage. One respondent without any social media experience and 64 individuals who did not use private (restricted) Instagram accounts were excluded, resulting in 636 eligible responses. To identify invalid responses, five attention check questions with explicit instructions were included. Responses that failed to comply with these instructions were excluded. Consequently, 508 valid responses (226 from female, 278 from male, and 4 from other participants) were retained for the final analysis.

4.2 Survey Results: Perceptions of the Symbolic Aspects of PPB

Respondents were shown sample images like that in Fig. 2 and asked whether they had ever posted photos edited to obscure personal identification on private Instagram accounts (including Stories). They were also asked to indicate their agreement with several evaluative statements about such posts.

Firstly, 45.08% (229 respondents) reported having posted edited photos. Although the degree of editing was not measured in detail, the findings suggest that nearly half



Fig. 2. Images shown to participants as examples of edited photos

of the participants had experience engaging in explicit forms of PPB through photo modifications.

Secondly, perceptions of the appropriateness of sharing such edited photos with all approved followers were explored. Two statements were used to assess the perceived effectiveness of this practice as a privacy measure: “I believe editing and posting photos should be done as a matter of etiquette to protect privacy,” and “I believe editing and posting photos is a highly effective method for protecting privacy.” A substantial majority agreed with both: 75.98% for the former (see Fig. 3) and 77.95% for the latter (see Fig. 4). These results reflect a strong awareness of the necessity and efficacy of such explicit PPBs.

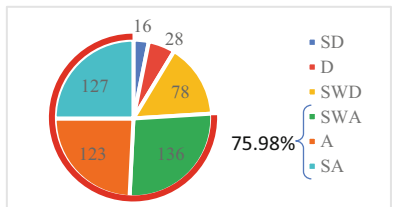


Fig. 3. Privacy Etiquette

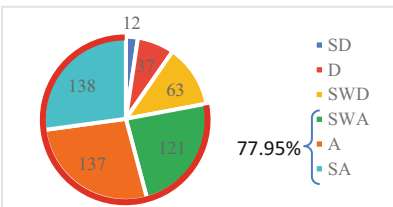


Fig. 4. Protective Effectiveness

The symbolic aspects of this behaviour were also examined using five additional evaluative statements. In response to “Posting edited photos gives the impression that one is overly concerned about privacy,” 43.90% agreed (Fig. 5). Furthermore, 40.94% (Fig. 6) agreed that “Posting edited photos makes it harder to post unedited ones.” Although these views were not dominant, they indicate that a significant proportion of respondents were conscious of how such actions may influence impression formation or subsequent behaviour.

Only 21.06% (Fig. 7) agreed with the statement “Posting edited photos will improve how others perceive me.” This, when considered alongside concerns about seeming excessively privacy-conscious, suggests that edited posts are more often seen as generating negative rather than positive social impressions. Nevertheless, agreement was relatively low for statements such as “Posting edited photos may make viewers feel uncomfortable” (22.05%, Fig. 8) and “Posting edited photos may make me appear distrustful of others” (21.46%, Fig. 9). These findings indicate that respondents did not generally believe that such behaviour would cause serious reputational harm.

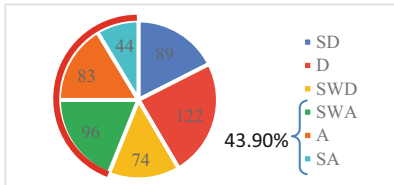


Fig. 5. Excess Privacy Impression

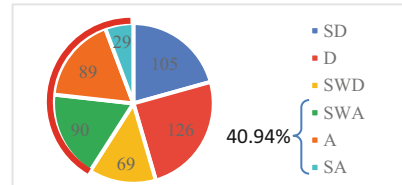


Fig. 6. Unedited Post Hindrance

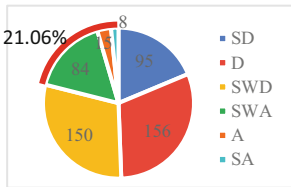


Fig. 7. Self-Image Improvement

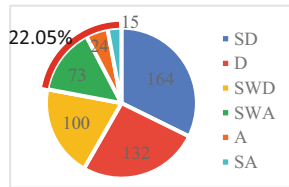


Fig. 8. Viewer Discomfort

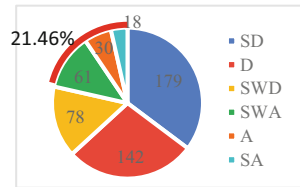


Fig. 9. Viewer Distrust

Taken together, the survey results point to two key insights. First, while explicit PPBs, such as editing photos before posting, are widely regarded as both necessary and effective, they may also raise concerns regarding their impact on self-presentation. This may lead to reluctance in adopting such behaviours. Second, although these practices are not typically seen as fundamentally damaging to one's image, they are perceived as somewhat awkward or socially fraught, thereby creating a subtle but significant barrier to their implementation.

4.3 Survey Results: Factors Influencing the Symbolic Aspects of PPB

Individuals' perceptions of the symbolic aspects of PPB are likely to be shaped by a variety of factors. This section focuses on two such influences: the perceived psychological distance from the audience to whom the content is addressed, and individuals' perceptions of privacy risk. Specifically, we examined whether these factors were associated with variation in how the symbolic dimensions of PPB were understood.

The symbolic interpretation of a person's actions is typically shaped by the nature of their relationship with the recipient [15]. For instance, when someone expresses an opinion in a forceful tone, the impression it creates may differ considerably depending on the audience's familiarity with the speaker. A close friend may interpret the behaviour as playful or good-humoured, whereas a more distant acquaintance might perceive it as evidence of a short temper or a domineering personality. In light of this, we extended the evaluation employed in the previous section (which focused on posts shared with all followers) by asking respondents to indicate their agreement or disagreement with the same evaluative statements in the context of posts shared exclusively with "Close Friends".

The distributions of agreement and disagreement for both recipient categories, along with the results of statistical tests for significant differences, are presented in Table 1.

For the two statements concerning the perceived necessity and effectiveness of posting edited photos as a form of PPB, statistically significant differences were observed at the 0.01% level. Specifically, respondents were less likely to regard such behaviour as necessary or effective when addressing close friends compared to a broader audience.

Table 1. Chi-square Analysis of Perceptions of Explicit PPB: All Followers vs. Close Friends

	Privacy Etiquette		Protective Effectiveness		Excess Privacy Impression		Unedited Post Hindrance	
	Disagree	Agree	Disagree	Agree	Disagree	Agree	Disagree	Agree
AF	122	386	112	396	285	223	300	208
CF	254	254	175	333	289	219	325	183
Statistics	$\chi^2(1) = 73.565^{***}$, $\phi = -0.269^{***}$		$\chi^2(1) = 19.274^{***}$, $\phi = -0.138^{***}$		$\chi^2(1) = 0.064$, $\phi = -0.008$		$\chi^2(1) = 2.598$, $\phi = -0.051$	
	Self-Image Improvement		Viewer Discomfort		Viewer Distrust			
	Disagree	Agree	Disagree	Agree	Disagree	Agree		
AF	401	107	396	112	399	109		
CF	401	107	398	110	341	167		
Statistics	$\chi^2(1) = 0.000$, $\phi = 0.000$		$\chi^2(1) = 0.023$, $\phi = -0.005$		$\chi^2(1) = 16.734^{***}$, $\phi = 0.128^{***}$			

* $p < .05$, ** $p < .01$, *** $p < .001$. All p-values are Bonferroni-adjusted.

In contrast, with regard to the symbolic aspects of PPB, no statistically significant differences were found in four of the five evaluative statements. The sole exception—significant at the 0.01% level—was the statement suggesting that posting edited photos might imply a lack of trust in the viewer. Here, agreement was significantly more frequent when the recipient group comprised close friends. This finding suggests that when content is shared with psychologically closer individuals, concerns may arise that such protective actions will be interpreted as a sign of distrust.

In summary, although the nature of the relationship with the audience does not substantially affect overall perceptions of the symbolic dimensions of PPB, it does influence specific aspects. Notably, when sharing with close friends, individuals appear more likely to question both the necessity and the effectiveness of privacy-protective behaviours, and to harbour concerns about the potential for negative symbolic interpretations—particularly those relating to trust. These tendencies may serve as disincentives to practising PPB in contexts involving close social connections, such as “Close Friends” sharing features on social media.

We also examined how individuals’ perceptions of privacy risk relate to their evaluations of PPB. Privacy risk perception was measured using two approaches. The first employed the Westin Privacy Index (WPI), which gauges general privacy consciousness based on responses to three items concerning: (1) perceived loss of control over personal data, (2) trust in third-party data handlers, and (3) confidence in data protection laws and regulations. Based on these responses, participants were categorised into one of three

groups: Privacy Unconcerned, Privacy Pragmatists, or Privacy Fundamentalists [16]. For instance, respondents who reported a loss of control over personal data but nonetheless expressed trust in both data handlers and regulatory frameworks were classified as Privacy Unconcerned. Those expressing distrust on both counts were labelled Privacy Fundamentalists, while the remainder were identified as Privacy Pragmatists.

Following this classification, the 508 valid respondents were distributed as follows: 131 Privacy Unconcerned, 250 Privacy Pragmatists, and 127 Privacy Fundamentalists. Chi-square tests were then conducted to determine whether group membership was associated with differences in response patterns to the evaluative statements discussed earlier. Frequency distributions and test statistics are reported in Table 2. No statistically significant differences were found among the groups for either of the two statements regarding the evaluation of PPB or the five statements relating to its symbolic dimensions. These results suggest that privacy consciousness, as operationalised through WPI classifications, does not significantly influence how individuals evaluate or interpret the symbolic implications of posting edited photos.

Table 2. Chi-square Analysis of Perceptions of Explicit PPB: Privacy Unconcerned, Privacy Pragmatists or Privacy Fundamentalists

	Privacy Etiquette		Protective Effectiveness		Excess Privacy Impression		Unedited Post Hindrance	
	Disagree	Agree	Disagree	Agree	Disagree	Agree	Disagree	Agree
Unconcerned	22	109	20	111	76	55	76	55
Pragmatists	65	185	61	189	137	113	149	101
Fundamentalists	35	92	31	96	72	55	75	52
Statistics	$\chi^2 (2) = 5.157$, $\phi = 0.101$		$\chi^2 (2) = 4.721$, $\phi = 0.096$		$\chi^2 (2) = 0.385$, $\phi = 0.028$		$\chi^2 (2) = 0.089$, $\phi = 0.013$	
	Self-Image Improvement		Viewer Discomfort		Viewer Distrust			
	Disagree	Agree	Disagree	Agree	Disagree	Agree		
Unconcerned	102	29	107	24	107	24		
Pragmatists	195	55	189	61	191	59		
Fundamentalists	104	23	100	27	101	26		
Statistics	$\chi^2 (2) = 0.889$, $\phi = 0.042$		$\chi^2 (2) = 1.910$, $\phi = 0.061$		$\chi^2 (2) = 1.519$, $\phi = 0.055$			

* $p < .05$, ** $p < .01$, *** $p < .001$. All p-values are Bonferroni-adjusted.

In addition to the WPI, we employed a more direct measure of privacy risk perception. Respondents were asked, “Do you feel at risk of a privacy breach when communicating via social media?” and were grouped according to their responses. The distributions of responses and the corresponding chi-square statistics are provided in Table 3. Among the seven evaluative statements, only one—“Posting edited photos is necessary as a matter of etiquette to protect privacy”—exhibited a statistically significant difference at the 0.1% level. Respondents who perceived privacy risks were more likely to agree

with this statement than those who did not. For all remaining statements, including those addressing symbolic aspects of PPB, no significant group differences were observed.

Table 3. Chi-square Analysis of Perceptions of Explicit PPB: No Risk Group and Feel Risk Group

	Privacy Etiquette		Protective Effectiveness		Excess Privacy Impression		Unedited Post Hindrance	
	Disagree	Agree	Disagree	Agree	Disagree	Agree	Disagree	Agree
No Risk	70	147	53	164	117	100	128	89
Feel Risk	52	239	59	232	168	123	172	119
Statistics	$\chi^2(1) = 14.103^{***}$, $\phi = 0.167^{***}$		$\chi^2(1) = 1.245$, $\phi = 0.050$		$\chi^2(1) = 0.735$, $\phi = -0.038$		$\chi^2(1) = 0.001$, $\phi = -0.001$	
	Self-Image Improvement		Viewer Discomfort		Viewer Distrust			
	Disagree	Agree	Disagree	Agree	Disagree	Agree		
No Risk	182	35	179	38	171	46		
Feel Risk	219	72	217	74	228	63		
Statistics	$\chi^2(1) = 5.547$, $\phi = 0.104$		$\chi^2(1) = 4.535$, $\phi = 0.094$		$\chi^2(1) = 0.015$, $\phi = 0.005$			

* $p < .05$, ** $p < .01$, *** $p < .001$. All p-values are Bonferroni-adjusted.

Taken together with the findings based on the WPI classifications, these results indicate that privacy risk perception does not meaningfully affect how individuals interpret the symbolic aspects of PPB. Regardless of whether individuals perceive themselves to be at risk, their evaluations of the symbolic consequences of protective behaviour, such as concerns about impression management, tend to align with the general patterns identified in the previous section.

5 Conclusions

This study explored the symbolic dimensions of explicit privacy-protective behaviours (PPBs) on private social media accounts, using data from semi-structured interviews and a questionnaire survey. The analysis showed that around 40% of participants believed these behaviours influenced how others perceived them on social media and shaped the communicative responses they received. Although participants did not view these effects as overtly negative, they still considered them undesirable and preferred to avoid them. These findings suggest that symbolic interpretations of explicit PPBs can act as subtle yet meaningful psychological barriers to their adoption. In this respect, the symbolic dimension may distort otherwise rational evaluations of PPBs and offers a useful theoretical lens through which to understand the privacy paradox more fully.

The study also found that symbolic awareness was largely independent of participants' general level of privacy concern. However, when participants perceived their audience as socially or psychologically close—as in personal relationships—they expressed greater concern that these behaviours might imply distrust and thus carry more negative connotations. While this study did not adopt an experimental design and cannot establish causality, it nevertheless points to important directions for future research.

These insights highlight the need to account for factors such as the explicitness and visibility of PPBs when designing legal frameworks, technological tools, and privacy-related services. Future studies will employ hypothesis-driven approaches to examine how PPB explicitness affects user behaviour and how this relationship may be moderated by contextual or cultural factors.

In the Japanese context, interpersonal communication often avoids explicit expression, driven by sensitivity to others' feelings and anticipated reactions. Consequently, overt assertions of privacy rights may be viewed as socially inappropriate [17]. Researchers must therefore consider these cultural dynamics when interpreting the symbolic significance of explicit PPBs. Cross-cultural comparative research represents a vital next step in advancing this line of inquiry.

Acknowledgments. This study was supported by the JSPS Grant-in-Aid for Scientific Research (Grant Number 23K01545 and 25K05673) and a grant from the Telecommunication Advancement Foundation of Japan. We would also like to appreciate Professor Hidenobu Sai (Ehime University), Associate Professor Takayuki Tanimoto (Ehime University), Associate Professor Miha Takubo (Ehime University) for their kind supports in collecting the survey responses, and all the participants in the survey.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Kröger, J.L., Lutz, O.H.M., Ullrich, S.: The myth of individual control: mapping the limitations of privacy self-management. *SSRN Electron. J.* 3881776. (2021)
2. Hull, G.: Successful failure: what foucault can teach us about privacy self-management in a world of Facebook and big data. *Ethics Inf. Technol.* **17**, 89–101 (2015)
3. Waldman, A.E.: Cognitive biases, dark patterns, and the 'privacy paradox.' *Curr. Opin. Psychol.* **31**, 105–109 (2020)
4. Young, A.L., Quan-Haase, A.: Privacy protection strategies on Facebook: the internet privacy paradox revisited. *Inf. Commun. Soc.* **16**(4), 479–500 (2013)
5. Baruh, L., Secinti, E., Cemalcilar, Z.: Online privacy concerns and privacy management: a meta-analytical review. *J. Commun.* **67**(1), 26–53 (2017)
6. Mousavi, R., Chen, R., Kim, D.J., Chen, K.: Effectiveness of privacy assurance mechanisms in users' privacy protection on social networking sites from the perspective of protection motivation theory. *Decis. Support Syst.* **135**, 113323 (2020)
7. Adhikari, K., Panda, R.K.: Users' information privacy concerns and privacy protection behaviors in social networks. *J. Glob. Mark.* **31**(2), 96–110 (2018)
8. Rogers, R.W.: A protection motivation theory of fear appeals and attitude change I. *J. Psychol.* **91**(1), 93–114 (1975)

9. Barth, S., De Jong, M.D.: The privacy paradox—investigating discrepancies between expressed privacy concerns and actual online behavior—a systematic literature review. *Telematics Inform.* **34**(7), 1038–1058 (2017)
10. Blumer, H.: *Symbolic Interactionism: Perspective and Method*. Prentice-Hall, Inc., Upper Saddle river (1969)
11. Goffman, E.: *The presentation of self in everyday life*. Doubleday & Company Inc. (1959)
12. Brake, D.R.: Who do they think they're talking to? Framings of the audience by social media users. *Int. J. Commun.* **6**, 1056–1076 (2012)
13. Hogan, B.: The presentation of self in the age of social media: distinguishing performances and exhibitions online. *Bull. Sci. Technol. Soc.* **30**(6), 377–386 (2010)
14. Goffman, E.: *Behavior in public places: Notes on the social organization of gatherings*. The Free Press of Glencoe (1963)
15. Adams, A.A., Murata, M., Orito, Y.: The Japanese sense of information privacy. *AI Soc.* **24**, 327–341 (2009)
16. Jai, T.M.C., King, N.J.: Privacy versus reward: do loyalty programs increase consumers' willingness to share personal information with third-party advertisers and data brokers? *J. Retail. Consum. Serv.* **28**, 296–303 (2016)
17. Orito, Y., Murata, K.: Privacy protection in Japan: cultural influence on the universal value. *Electron. Proc. Ethicomp*, 1–9 (2005)



Operational Archives and the Right to Be Forgotten

Klara Källström^(✉)

Department of Applied Information Technology, University of Gothenburg, Gothenburg, Sweden

`klara.kallstrom@gu.se`

Abstract. This paper explores how algorithmic systems reconfigure established notions of the image, reframing it as fragmented, actionable data units rather than a coherent representational form, thereby raising critical questions about privacy rights such as the Right to Be Forgotten. Drawing on the example of the Swedish police's use of facial recognition technology, it illustrates how photographs are reconstituted as actionable elements within predictive and classificatory infrastructures. As archives become computational and machine-driven, established understandings of deletion, autonomy, and accountability are increasingly destabilised. The analysis shows that the conceptual and regulatory limits of erasure become particularly salient in contexts where data is no longer merely stored but persistently circulated and acted upon. The paper concludes that legal frameworks premised on coherent data subjects and discrete visual artefacts are inadequate to the distributed logics of algorithmic governance, and it calls for a reimagining of archival and data governance practices that recognise the enduring operational life of images beyond their initial capture.

Keywords: Algorithmic governance · Operational images · Informational Privacy · Right to Be Forgotten · Archival Practices · Data Infrastructures · Facial Recognition · Ethics

1 Introduction

Contemporary forms of algorithmic governance increasingly challenge established assumptions about autonomy, particularly in relation to how personal data such as images are processed, circulated, and acted upon. Within public administration, the integration of algorithmic systems has transformed the role of visual material: images, once conceived as stable representations intended for human interpretation, are now fragmented, recomposed, and operationalised as data points within computational infrastructures of classification and control [1]. This ontological shift marks the industrialisation of images, whereby they cease to function primarily as visual artefacts and instead operate as dynamic components within automated decision-making systems [17, 30].

This paper investigates the ontological and epistemological shifts in the role of images within algorithmic systems, using the Swedish police's use of Clearview AI's

facial recognition technology as an illustrative example. It examines how images, once absorbed into computational infrastructures, are reconstituted as dispersed, machine-readable entities. This transformation challenges the legal and technical viability of regulatory frameworks such as the Right to Be Forgotten, which grants individuals the right to request the erasure of personal data that is no longer necessary for the purposes for which it was collected or that has become objectionable, under certain conditions [23]. When visual material becomes embedded within predictive and classificatory architectures, its withdrawal exposes the already unstable foundations of archival systems, where erasure is not only technically complex but also historically contingent [6].

The paper reflects on the broader conditions under which visual data persists within algorithmic governance. Rather than functioning as static records, images operate as distributed signals shaped by institutional power and data flows [6, 12, 22]. These shifts give rise to a central research question:

How does the operationalisation of images shape the ability to uphold the Right to Be Forgotten in archival practices?

Assuming that archives are no longer merely repositories but dynamic infrastructures entangled with algorithmic processes of extraction, recomposition, and circulation, archival practices emerge less as mechanisms for the storage and retrieval of static records and more as integral components of computational governance [23]. When understood as operational rather than representational, the archive necessitates new theoretical and regulatory frameworks capable of addressing how memory, visibility, and erasure are enacted within algorithmic regimes [1].

2 Theoretical Foundation

Informational privacy is often framed as the individual's capacity to dissociate from persistent digital traces [19, 33, 38]. This framing assumes a coherent, autonomous subject capable of controlling personal data. Among numerous critiques of the autonomous subject, Gilles Deleuze's concept of the *dividual* offers a particularly incisive challenge, proposing that identity within computational systems is fragmented and operationalised through data points, behaviours, and metrics [11]. Within this framework, privacy concerns shift away from the ownership of discrete data toward the aggregation and circulation of informational fragments used for classification and prediction [2, 7, 27]. A similar ontological shift characterises the treatment of images: once processed algorithmically, images are transformed into dispersed data points embedded within computational infrastructures [12, 29]. Like personal data, the image becomes fragmented, recontextualised, and ultimately inaccessible as a coherent object.

As algorithmic systems increasingly shape how data is processed within governance frameworks, the inherent instability of autonomy—central to rights such as the Right to Be Forgotten (Art. 17 GDPR)—comes to the fore [3, 31]. While this right seeks to mitigate the persistence of digital traces, it rests on the assumption that deletion protects privacy. Yet this premise falters in distributed and predictive infrastructures, where data, visual or otherwise, is continuously replicated, recombined, and embedded within classificatory systems. Deletion becomes not a definitive act but a contingent intervention shaped by archival logics and infrastructural resistance [6, 23, 28].

The landmark 2014 decision in *Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* further complicated the implementation of the Right to Be Forgotten. It effectively positioned Google and other search engines as de facto arbiters of data erasure, shifting the burden of enforcement onto private corporations (Court of Justice of the European Union [CJEU], 2014; Kuczerawy, 2015; Ausloos, 2020). This delegation of regulatory power illustrates the legal ambiguity surrounding RTBF: while conceived as a user-centric right, it has paradoxically consolidated the role of dominant platforms in governing data visibility and erasure—raising questions about transparency, accountability, and structural power in digital privacy governance.

Understanding these tensions requires moving beyond individual rights claims toward a broader analysis of the infrastructural and archival systems through which data governance now operates. This shift demands critical engagement with algorithmic infrastructures that shape behaviour through data flows [7, 22], rooted in a genealogy of classificatory and surveillance practices shaped by historically contingent regimes of power and knowledge [20]. Archival systems have long operated as instruments of institutional power, structuring access, memory, and control [6], with photography playing a pivotal role in biometric surveillance and criminological classification [5, 27, 35]—logics that continue to inform contemporary facial recognition technologies [3, 8].

Blaschke et al.'s [5] reflections on the fractured ontology of photographs under generative AI offer a productive framework for understanding these developments. In systems like Clearview AI, photographs scraped from social media and public sources are no longer treated as documentary records but as fragmented data points instrumentalized for biometric computation [26]. Much like training datasets in generative AI models, these images are operationalised as they are parsed into machinic patterns, recomposed within predictive recognition infrastructures, and severed from their original contexts. This reconfiguration prompts a reconsideration of the photograph's epistemological status. No longer perceived as evidentiary traces anchored in human memory, they emerge instead as volatile informational materials circulating through opaque computational networks. Clearview AI thus exemplifies how contemporary data practices transform visual archives into dynamic operational systems, where the boundaries between memory, identity, and synthetic reproduction become increasingly blurred.

The shift from physical archives to expansive digital repositories, and more recently to algorithmically driven infrastructures, has intensified these dynamics. Unbound by spatial constraints, algorithmic systems enable indefinite data retention, recombination, and circulation, complicating regulation around deletion, access, and classification [7, 29]. The contemporary emergence of generative AI systems adds a further layer of complexity. Visual material is no longer merely stored or sorted but becomes the raw substrate for synthetic recombination. Images, fragmented into operational data, are extracted, parsed, and remobilised within training datasets designed to produce new outputs [5].

This reconfiguration challenges established archival frameworks that assume records are stable and traceable entities. Instead, under generative paradigms, informational traces persist in unpredictable forms and are continually recombined in ways that blur distinctions between original and derivative, human and machinic [34]. Regulatory

strategies premised on managing discrete, identifiable data increasingly falter in the face of systems where data is not simply retained but continuously enacted and repurposed across distributed networks [8]. These challenges are compounded by the opacity of algorithmic infrastructures, which, as Paglen [29] explains, increasingly produce invisible images—visual data generated by and for machines that function outside the scope of human perception or interpretability. The underlying processes of extraction, transformation, and redistribution often remain inaccessible to scrutiny.

What emerges from these developments are not only technical but also epistemic and ontological challenges to contemporary regulatory frameworks. While rights such as the GDPR's Right to Be Forgotten and the Law Enforcement Directive (LED)'s protections for law enforcement data attempt to address the persistence of personal information, they remain fundamentally tethered to a conception of the subject as discrete, coherent, and sovereign. However, as algorithmic infrastructures increasingly operate on fragmented and distributed data flows in which individuals are rendered as dividual composites of behavioural traces and predictive scores, these legal instruments encounter profound ontological limits. Rights such as access, deletion, or rectification are premised on the traceability and singularity of the data subject, assumptions that falter in contexts where data is no longer anchored to stable identities but circulates through operational and machinic systems beyond direct human agency.

Thus, contemporary governance frameworks must confront not only the opacity and persistence of algorithmic archives but also the fundamentally altered ontology of visual and personal data itself. Infrastructures of surveillance, prediction, and synthetic generation are sustained not merely by the retention of records, but by the perpetual remaking of informational life beyond the bounds of human agency or direct institutional control.

To explore how tensions between data subjectivity, infrastructural opacity, and regulatory governance manifest in practice, this study examines the Swedish Police Authority's use of Clearview AI's facial recognition system as an illustrative example. Although the technology was not officially sanctioned by the authority, it was nonetheless used by individual officers, raising critical questions about institutional oversight and accountability. The analysis demonstrates how this unsanctioned use exemplifies deeper structural tensions in the governance of visual data within algorithmic systems.

3 The Swedish Police's Use of Unauthorised Facial Recognition

Clearview AI's platform, built on a database of over three billion images scraped without consent, exemplifies the transformation of images into biometric data within algorithmic surveillance infrastructures [32, 36]. The controversy intensified in 2020 when *BuzzFeed News* leaked Clearview's client list, revealing its use by law enforcement agencies across the United States, Canada, and several European countries, including Sweden. In Canada, Finland, and Sweden, officers accessed the platform without institutional approval, blurring the boundaries between state oversight and commercial infrastructures [1, 13, 14, 36].

In 2021, the Swedish Authority for Privacy Protection (IMY) concluded that the Police Authority had violated the Swedish Criminal Data Act by using Clearview AI

without a required data protection impact assessment (Dnr 4756-21). The ruling underscores the challenges of enforcing protections such as the Right to Be Forgotten in a context where images are embedded within distributed systems that resist deletion [28]. IMY ordered that affected individuals be notified and that the unlawfully processed data be erased [13, 23].

Importantly, data processing by law enforcement agencies such as the Swedish Police falls under the scope of Directive (EU) 2016/680—commonly referred to as the Law Enforcement Directive (LED)—which specifically governs the processing of personal data by competent authorities for the purposes of criminal justice, such as prevention, investigation, detection, and prosecution of offences. While both the LED and the GDPR share foundational principles, the LED introduces more specific criteria for data retention, deletion, and impact assessments in policing contexts. Notably, Article 16 of the LED limits the applicability of erasure rights when data is processed for criminal justice purposes, allowing for exceptions related to ongoing investigations and public interest. This legal distinction underscores the complexity of applying RTBF to facial recognition systems within law enforcement, where privacy protections are conditional and context-dependent [16].

This paper uses the Swedish Police Authority's use of Clearview AI's facial recognition technology as an illustrative example, drawing on and referring to previous empirical research that has investigated the case in detail [13, 14]. Specifically, these materials include public records from the Swedish Authority for Privacy Protection (IMY), the Police Authority, and the Administrative Court in Stockholm, obtained under Sweden's Principle of Public Access to Information, as outlined in the Freedom of the Press Act (1949:105).

This closer examination of the Clearview AI case brings into focus persistent tensions between legal frameworks and the infrastructural realities of data persistence, fragmentation, and circulation. Across administrative rulings and communications, efforts to regulate facial recognition technologies appear constrained by the structural limits of distributed algorithmic systems. These observations inform the paper's broader theoretical argument that the Right to Be Forgotten, within contemporary algorithmic archives, must contend not only with legal and institutional barriers but also with epistemic and material shifts in the ontology of data (Table 1).

4 The Ontological Challenge of the Right to Be Forgotten

This discussion of the Swedish police's use of Clearview AI's facial recognition technology highlights the difficulty of upholding individual autonomy in the face of data operationalisation. As state institutions integrate proprietary algorithmic infrastructures into their operations, images and personal data alike are transformed into operational components—no longer discrete, retrievable records, but dynamic inputs for classification, recognition, and prediction [1]. Through the example of Clearview AI and an examination of legal decisions and administrative correspondence, it becomes evident that the technology was used by individual officers outside formal approval channels, resulting in a breakdown of institutional oversight and legal accountability.

Table 1. Overview of the collected documents

Authority	Type of document
The Swedish Authority for Privacy Protection (IMY)	Investigation into the use of Clearview AI (Dnr DI-2020–2719, date: 2020–03-05) Request for supplementation (Dnr DI-2020–2719, date: 2020–03-30) Decision after the inspection (Dnr DI-2020–2719, date: 2021–02-10)
The Swedish Police Authority	Investigation into the use of Clearview AI (Dnr A126.614/2020, date: 2020–03-19) Request for supplementation (Dnr A126.614/2020, date: 2020–05-07) Appeal regarding the IMY’s decision (Dnr A126.614/2020, date: 2021–03-01) Completion of previously filed appeal regarding the IMY’s decision (Dnr A126.614/2020, date: 2021–03-05)
The Administrative Court in Stockholm	Decision of the Administrative Court (Dnr 4756–21, date: 2021–09-30)
The Court of Appeal in Stockholm	Decision of the Court of Appeal (Dnr 7678–21, date: 2022–11-07)

This shift challenges the legal and conceptual foundations of rights such as the Right to Be Forgotten. The issue is not merely whether data can be deleted, but whether withdrawal remains meaningful once data is mobilised within computational infrastructures. Once processed, images are fragmented, distributed, and rendered invisible—not merely stored but used, and deeply entangled with material, geopolitical, and logistical networks of appropriation and circulation [9]. In this context, the image becomes part of a global supply chain of data industrialisation, to be mined, reformulated, and propagated. In this Farockian sense, archives cease to function as static repositories and instead become operationalised systems of capture and recomposition, where images are not preserved but enacted upon [1, 6, 17, 30].

Moreover, the RTBF is not an absolute right [23]. As stated in Recital 4 of the GDPR, the protection of personal data must be balanced against other fundamental rights and freedoms, including freedom of expression and access to information. These tensions become especially pressing in cases involving biometric data in public interest contexts. The Swedish example illustrates how RTBF claims intersect with competing obligations, such as the right to informational self-determination on one hand and institutional responsibilities for investigation, documentation, and security on the other. Understanding the RTBF as a principle of legal balancing, rather than a clear-cut entitlement, is essential for appreciating its limits within algorithmic governance.

Jussi Parikka’s theorisation of operational images extends Farocki’s insight by implicitly framing the archive not simply as a site of storage, but as a dynamic, machinic

infrastructure of continuous calculation and circulation [30]. For Parikka, such infrastructures constitute the technical conditions of contemporary vision and governance, in which images function less as representations than as inputs for automated action. This perspective is particularly useful in analysing the example of the Swedish police's use of Clearview AI, an infrastructure that transforms facial images into operational data streams, enabling continuous biometric recognition beyond the reach of traditional archival or legal mechanisms. Parikka's work clarifies how algorithmic image systems displace conventional notions of memory, deletion, and retrieval, demanding new frameworks for understanding how data persists, acts, and governs.

Michelle Henning's notion of "the unfettered image" offers a complementary conceptual lens. She describes the image as simultaneously fixed and migratory: it is anchored in visual form, yet constantly moving as an object through screens, infrastructures, and technical systems [21]. This paradox captures the ontological condition of images in systems like Clearview AI. Such images retain a recognisable visuality, but are parsed and mobilised as modular units within predictive infrastructures. Their unfettered nature means they can no longer be clearly tied to an origin, an owner, or even a stable context. They circulate far beyond their institutional beginnings and outside the bounds of representational intent. This compounds the difficulty of withdrawal: how does one erase an image that no longer behaves as an image, but as a dispersed signal or algorithmic function?

In such a paradigm, deletion becomes a technically and epistemologically unstable intervention. Legal mechanisms premised on the coherent data subject and the reversibility of processing collapse under infrastructures that retain, recombine, and act on data without fixed boundaries. The concept of autonomy proves increasingly untenable in environments where individuals are parsed into dividual fragments, legible only through patterns and predictions. Rather than enabling control, autonomy is revealed as a fragile construct, outpaced by systems that operate beyond the grasp of intentionality.

The findings of the Swedish Authority for Privacy Protection (IMY) further underscore this disjunction. The directive to delete and notify presumes a subject who can be located, addressed, and restored to privacy. These assumptions prove incompatible with algorithmic infrastructures in which data persists not as static content but as actionable process. In the example of the Swedish police's use of Clearview AI, as documented in IMY's investigation and court records, attempts to enforce deletion encountered both practical and epistemic barriers, such as the inability to trace where facial images had been processed, stored, or transmitted. This example demonstrates that erasure can no longer be conceptualised as the removal of discrete items from storage. Instead, if it is possible at all, erasure must contend with a broader ecology of circulation, recomposition, and operationalisation.

The industrialisation of data thus demands not only structural adjustments but a rethinking of the foundations of privacy governance—one that accounts for the performative and distributed life of data beyond the moment of capture. A revised framework must grapple with the realities of the operational archive and the unfettered image: visual data that moves, acts, and persists across jurisdictions and technical regimes, without ever fully coming to rest [1].

This legal landscape is further complicated by the EU Artificial Intelligence Act, which entered into force on 1 August 2024 [16]. The Act introduces a risk-based regulatory framework for AI systems, with its provisions being implemented progressively. Notably, Article 5 prohibits certain uses of real-time remote biometric identification in publicly accessible spaces. However, it includes broad exceptions for law enforcement, national security, and the prevention of serious crime, provided specific safeguards are met. Although the Act is now being implemented [13], the application of these prohibitions and their exceptions underscores the fragility of such regulatory measures: even where legal restrictions exist, the scope of derogations risks undermining their effectiveness [37]. Any robust analysis of the Right to Be Forgotten in algorithmic contexts must therefore be situated within this shifting legal terrain, attentive to how proposed safeguards may be compromised by institutional exceptions and asymmetries of power.

References

1. Amore, L.: *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Duke University Press (2020)
2. Amore, L., Campolo, A., Jacobsen, B., Rella, L.: A world model: on the political logics of generative AI. *Polit. Geogr.* **113**, 103134 (2024). <https://doi.org/10.1016/j.polgeo.2024.103134>
3. Ananny, M., Crawford, K.: Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc.* **20**(3), 973–989 (2018). <https://doi.org/10.1177/1461444816676645>
4. Ausloos, J.: *The Right to Erasure in EU Data Protection Law: From Individual Rights to Effective Protection*. Oxford University Press (2020). <https://doi.org/10.1093/oso/9780198847977.001.0001>
5. Blaschke, E., Bonhomme, M., Joschke, C., Somaini, A.: Introduction: photographs and algorithms. *Transborder* **9** (2025). <https://doi.org/10.4000/13dwo>
6. Bowker, G.C.: *Memory Practices in the sciences*. MIT Press, Cambridge (2005)
7. Bowker, G.C., Star, S.L.: *Sorting Things Out: Classification and its Consequences*. MIT Press, Cambridge (1999)
8. Crawford, K.: *Atlas of AI: Power, Politics, and the Planetary costs Of Artificial Intelligence*. Yale University Press, Yale (2021). <https://doi.org/10.12987/9780300252392>
9. Crawford, K., Joler, V.: *Anatomy of an AI system: The Amazon Echo as an anatomical map of human labor, data, and planetary resources*. Share Lab & The AI Now Institute (2018). <https://anatomyof.ai>
10. Court of Justice of the European Union (CJEU). Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González, Case C-131/12, ECLI:EU:C:2014:317 (2014). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62012CJ0131>
11. Deleuze, G.: Postscript on the societies of control. *October* **59**, 3–7 (1992)
12. Dewdney, A., Sluis, K. (eds.): *The Networked Image in Post-Digital Culture*. Routledge, Abingdon (2023)
13. Eneman, M., Ljungberg, J.: AI and governance dilemmas for law enforcement agencies. In: Iveroth, E., Lindvall, J., Magnusson, J. (eds.) *Leading Digital Transformation: Management, Governance and Control*. Routledge (2025)
14. Eneman, M., Ljungberg, J., Raviola, E., Rolandsson, B.: The sensitive nature of facial recognition: tensions between the Swedish police and regulatory authorities. *Inf. Polity* **27**(2), 219–232 (2022). <https://doi.org/10.3233/IP-211538>

15. European Union. Directive (EU) 2016/680 on data processing in criminal justice. Official Journal of the European Union, L 119/89 (2016). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016L0680>
16. European Union. Regulation (EU) 2024/1352 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Official Journal of the European Union, L 202, 1–127 (2024). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1352>
17. Farocki, H.: Phantom images. *Public* **29**, 12–22 (2004)
18. Farocki, H.: Eye/Machine I–III [Video essay series]. Harun Farocki Filmproduktion (2001–2003)
19. Floridi, L.: *The Onlife Manifesto: Being Human in a Hyperconnected Era*. Springer, Heidelberg (2015). <https://doi.org/10.1007/978-3-319-04093-6>
20. Foucault, M.: *Discipline and punish: The birth of the prison* (A. Sheridan, Trans.). Pantheon Books (1977)
21. Henning, M.: *Photography: The Unfettered Image*. Routledge, Abingdon (2022)
22. Jasanoff, S.: *States of Knowledge: The co-Production of Science and the Social Order*. Routledge, Abingdon (2004). <https://doi.org/10.4324/9780203413845>
23. Kosta, E.: Algorithmic state surveillance: challenging the notion of agency in human rights. *Regul. Gover.* **16**(1), 212–224 (2022). <https://doi.org/10.1111/regg.12331>
24. Kuczerawy, A.: Intermediary liability & freedom of expression: recent developments in the EU notice & action initiative. *Comput. Law Secur. Rev.* **31**(1), 46–56 (2015). <https://doi.org/10.1016/j.clsr.2014.11.004>
25. Latour, B.: *Reassembling the Social: An Introduction to Actor-Network Theory*. Oxford University Press, Cambridge (2005)
26. KU Leuven Centre for IT & IP Law. Clearview AI: Illegally collecting and selling our faces in total impunity – Part I (2022). <https://www.law.kuleuven.be/citip/blog/clearview-ai-illegally-collecting-and-selling-our-faces-in-total-impunity-part-i/>
27. Lyon, D.: *Surveillance Studies: An Overview*. Polity Press (2014)
28. Mantelero, A.: The EU proposal for a general data protection regulation and the roots of the ‘Right to Be Forgotten.’ *Comput. Law Secur. Rev.* **29**(3), 229–235 (2013). <https://doi.org/10.1016/j.clsr.2013.03.010>
29. Paglen, T.: *Invisible images (Your pictures are looking at you)*. The New Inquiry (2016). <https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/>
30. Parikka, J.: *Operational Images: From the Visual to the Invisual*. University of Minnesota Press (2023)
31. Pasquale, F.: *The Black Box Society: The Secret Algorithms that Control Money and Information*. Harvard University Press (2015). <https://doi.org/10.4159/harvard.9780674736061>
32. Rezende, I.N.: Facial recognition in police hands: assessing the ‘Clearview case’ from a European perspective. *New J. Eur. Crim. Law* **11**(3), 375–389 (2020). <https://doi.org/10.1177/2032284420948161>
33. Richards, N.M.: *Why Privacy Matters*. Oxford University Press, Cambridge (2022)
34. Rouvroy, A., Berns, T.: Algorithmic governmentality and prospects of emancipation: Disparateness as a precondition for individuation. (L. Carey-Libbrecht, Trans.). *Réseaux*, **177**(1), 163–196 (2013). <https://shs.cairn.info/journal-reseaux-2013-1-page-163?lang=en>
35. Sekula, A.: The body and the archive. *October* **39**, 3–64 (1986). <https://doi.org/10.2307/778312>
36. Shepherd, T.: The canadian clearview AI investigation as a call for digital policy literacy. *Surveill. Soc.* **22**(2), 179–191 (2024). <https://doi.org/10.24908/ss.v22i2.16300>

37. Veale, M., Borgesius, F.Z.: Demystifying the draft EU artificial intelligence act. *Comput. Law Rev. Int.* **22**(4), 97–112 (2021). <https://doi.org/10.9785/cr-2021-220402>
38. Véliz, C.: *Privacy is Power: Why and How You Should Take Back Control of Your Data*. Bantam Press (2020)



Smart Doorbells in a Surveillance Society

Anjela Mikhaylova^{1,2}(✉) , Sara Wilford¹ , Bernd Stahl² ,
and Laurence Brooks³

¹ De Montfort University, Gateway House, Leicester LE1 9BH, UK
{anjela.mikhaylova, sara}@dmu.ac.uk

² Computer Science, University of Nottingham, Nottingham NG8 1BB, UK
bernd.stahl@nottingham.ac.uk

³ University of Sheffield, Information School, The Wave, Sheffield S10 2AH, UK
l.brooks@sheffield.ac.uk

Abstract. Smart doorbells, promoted as tools of convenience and peace of mind, extend far beyond their intended role as home security enhancements. They introduce new forms of domestic surveillance by capturing audio-visual data around properties, blurring the boundaries between private and public space.

This paper presents findings from the first phase of a two-stage qualitative study exploring how stakeholders perceive and navigate the ethical, legal, and social implications of smart doorbell surveillance.

Drawing on a pilot survey conducted in a Leicester (UK) neighbourhood, the study identifies several emergent themes: shifts in community trust, the use of surveillance in neighbour conflicts, issues of incidental recording and consent, and growing discomfort with the normalisation of surveillance. These findings are analysed through the theoretical lenses of Foucault's *Panopticism* and Zuboff's *Surveillance Capitalism*, highlighting how smart doorbells reshape behaviour, visibility, autonomy, and power dynamics at the residential threshold.

In response to these concerns, the discussion incorporates normative frameworks, specifically *Privacy by Design*, *Ethics by Design*, and *Responsible Research and Innovation (RRI)*, to outline practical pathways for more responsible and ethical technology design and use.

Keywords: Smart doorbells · Surveillance Society · Panopticism · Ethics · Privacy

1 Introduction

Smart doorbell technology has transformed the very concept of the household front door over the past decade. Once passive fixtures used solely to signal a visitor's arrival, doorbells have evolved into staple components of residential surveillance sites within smart home ecosystems. This increasingly common presence of smart doorbells has been fuelled by the push by smart doorbell manufacturers such as Amazon (*Ring*), Google (*Nest*), and Anker (*Eufy*), to promote the idea of increased security and a sense of safety, peace of mind, and convenience. Equipped with motion sensors, high-definition cameras,

and microphones, smart doorbells enable users, as the tagline goes, to “remotely monitor, track, and record activities around their property boundaries, as well as interact in real time with individuals at their front doors” via apps and voice-activated assistants. The home conveys the sense of being under constant surveillance.

Yet beneath the reassuring language of security and convenience lies a more unsettling reality: smart doorbell technology is redefining norms of visibility, control, and privacy at the residential threshold, blurring the line between safety and continuous surveillance. Preliminary findings from the present study’s pilot survey indicate that their use gives rise to a spectrum of ethical issues related to surveillance, privacy, and data governance, as the audio and video recordings often occur without the knowledge or consent of those captured, including “*consumers and other individuals*” [32].

One key insight from the pilot is the further erosion of boundaries between public and private domains, particularly when the surveillance reach of these devices encroaches on shared spaces such as pavements, driveways, and shared entrances, thereby institutionalising private surveillance [1]. Building on this, the findings suggest that behind the veneer of protection, smart doorbells expand the reach of surveillance, transforming users into participants in a broader surveillance apparatus. What has been described as “*consumer-initiated surveillance*” [19] reflects a shift of power dynamics with private citizens, now empowered to surveil others without consent, becoming data collectors in public and semi-public spaces.

Furthermore, the pilot study suggests that beyond the surface-level advantages, smart doorbells introduce deeper questions regarding the fine line between security and extent of privacy intrusion and who controls access to the collected data, as the data are typically stored either locally on the device or uploaded to cloud platforms, where they may be retained, processed, or shared, often beyond the direct control of the user. These preliminary insights reveal that the ethical dilemmas surrounding smart doorbell use are not merely theoretical but are experienced by stakeholders in real-world settings.

Given these complexities, this research is guided by the following question: *In what ways do stakeholders perceive and balance the ethical concerns and social benefits of smart doorbell technology?*

The **aim** of this study is to explore the perspectives of stakeholders on the ethical and social implications of integrating smart doorbells into residential security. Drawing on pilot survey data, it examines the trade-offs between perceived security benefits, convenience, and privacy intrusions, alongside broader issues such as creeping expansion of surveillance, crime prevention, and public safety. In doing so, the research also considers unintended consequences of this technology, including its impact on social trust, community dynamics, and individual autonomy. By engaging with these themes, this research enriches the dialogue on responsible technology use and the evolving landscape of a surveillance-driven society.

The remainder of this paper is structured as follows: Sect. 2 reviews academic and non-academic literature on smart doorbells, privacy, and the social impact of surveillance dynamics. Section 3 presents the theoretical approach, focusing on Panopticism and Surveillance Capitalism. Section 4 details the two-phase methodology and pilot survey design. Section 5 presents pilot study key findings on how smart doorbells shape trust, behaviour, and privacy. Section 6 analyses these findings through theoretical and

normative lenses, proposing actionable responses informed by Privacy by Design, Ethics by Design, and Responsible Research and Innovation (RRI). Finally, Sect. 7 concludes this paper and outlines contributions, limitations, and avenues for future research.

2 Literature Review

This literature review adopts a dual-pronged approach, drawing on both peer-reviewed academic sources to ground the discussion in established theoretical and empirical research and non-academic literature, including legal cases, privacy policies, media reports, and investigative journalism, to contextualise real-world developments, public discourse, and regulatory challenges. Together, these sources help frame the emerging ethical and social questions surrounding smart doorbell surveillance in everyday settings.

Much of the existing literature subsumes these devices under broader discussions of smart home ecosystems [9, 16], inadvertently obscuring the distinct social and surveillance dynamics introduced at the residential threshold. Unlike indoor surveillance devices intended for private use, smart doorbells operate at the interface between private and public space. They routinely record individuals outside the home, including neighbours, delivery personnel, caretakers, and passersby, without their knowledge or consent. This raises questions about the erosion of privacy and the normalisation of ambient surveillance.

Recent academic research has begun to highlight the complex privacy challenges posed by these technologies [1, 18, 19, 22]. For example, Alshehri et al. [1] and Kelly [19] examine concerns from both homeowners' and bystanders' perspectives, revealing tensions around "*consumer-initiated surveillance*" and asymmetrical control over data. Furthermore, Selinger and Durant [33] argue that smart doorbells represent a form of state surveillance embedded into civilian life through partnerships with law enforcement, thereby raising concerns about due process and regulatory oversight. In addition, Grauer [18] draws attention to the underacknowledged role of audio recordings, noting that many users remain unaware these devices capture sound well beyond property boundaries, a phenomenon Schleusener [32] terms a "*sweeping surveillance nexus*" [19, 32].

Research has further linked smart doorbell use to discriminatory surveillance practices and the disproportionate targeting of marginalised communities [20, 36]. Kurwa [20] and Wong et al. [36] explore how smart surveillance technologies disproportionately affect racialised and low-income communities, often reinforcing patterns of biased reporting and observation. These findings raise further ethical concerns around who has the power to surveil, who is surveilled, and what social assumptions underpin these practices.

Beyond academia, legal commentary and high-profile cases have amplified concerns over the real-world consequences of domestic surveillance.

A Legal Precedent: Fairhurst v. Woodard. Legal precedent has begun to respond. A landmark UK case, *Fairhurst v. Woodard* (2021) [15], demonstrated that domestic surveillance systems, such as smart doorbells and floodlight cameras, that capture footage beyond a user's property can violate data protection laws. In this case, a judge found multiple breaches of the GDPR, including unlawful processing, lack of transparency, and excessive data retention. The use of multiple Ring cameras was ruled to have unjustifiably

invaded a neighbour's privacy, constituting harassment, and resulting in a £100,000 fine [15], with the affected neighbour forced to relocate due to a persistent sense of being watched. This precedent underscores that even privately owned surveillance tools must comply with data protection principles when their operation encroaches on public or shared spaces.

Smart doorbells have also played a key role in supporting criminal investigations. In Bedfordshire, UK, a video doorbell owner provided Ring footage to police that helped police identify a burglary suspect. The suspect's face and fingerprints were subsequently linked to a series of burglaries in the neighbourhood, leading to their arrest. Similarly, video doorbell footage played a critical role in the investigation of Sarah Everard's disappearance by assisting investigators in establishing Sarah's key locations and reconstructing the sequence of events leading up to her abduction, which contributed to the arrest of a police officer on suspicion of murder [35].

However, while cooperation with law enforcement may improve public safety, it has also prompted criticism. Some argue that involving residents in tasks that could be classified as police operations and reserved for trained officers may represent an attempt to cut policing costs or to outsource responsibilities to civilians. This dynamic blurs boundaries between state authority and private surveillance and contributes to the potential normalisation of informal civilian surveillance.

In a similar vein, Morris [24] critiques the integration of smart doorbells with law enforcement by highlighting the lack of legal safeguards and warning that the absence of warrants for accessing footage risks undermining civil liberties. He refers to this as "*tech exceptionalism*," a regulatory blind spot that creates conditions for pervasive surveillance and potential misuse. Selinger and Durant [33] further warn that such surveillance represents a "*slippery slope*," especially when platforms like Amazon Ring partner with police departments and facilitate civilian-to-state surveillance pipelines.

The Emergence of Biometric Surveillance. A further complication stems from the integration of biometric technologies into smart doorbells. Increasingly, these devices incorporate features such as facial recognition and voice analysis. These systems process immutable physiological or behavioural characteristics, including facial structures, gait, and vocal tone. The use of such biometric data introduces new risks: individuals may be persistently identified, profiled, and tracked, raising significant concerns about loss of anonymity and the emergence of lifetime surveillance [14].

The legal grey zones surrounding device placement, biometric data use, and informal data sharing with police contribute to an environment in which domestic surveillance may operate with little transparency or accountability despite its broad social impact. These findings underscore the urgent need for updated regulatory frameworks that go beyond general data protection principles. Clearer rules, enforceable standards, and public education are essential to bridging the gap between legal intent and on-the-ground practice.

3 Theoretical Framework

To better understand the ethical and social implications identified in the existing literature and supported by early findings from our pilot survey, this research applies two established theoretical frameworks: Foucault's concept of *Panopticism* and Zuboff's

theory of *Surveillance Capitalism* [17, 37]. These frameworks explain how surveillance restructures power, autonomy, and behaviour and provide a lens through which to analyse the role of smart doorbells in contemporary surveillance society. These frameworks not only guide the interpretation of stakeholder experiences but also shape the design and analytical focus of the study.

3.1 Monetising the Front Door: Surveillance Capitalism in Practice

In *Domestic Fortress: Fear and the New Home Front*, Atkinson (2016) examines evolving attitudes toward home security over the past several decades. Increasingly, domestic spaces are fortified using protective technologies, often echoing the logic of mediaeval architectural defence (pp. 107–131, pp. 175–193). Atkinson questions the underlying social and economic forces that have fuelled and cultivated feelings of heightened insecurity, fear, and anticipation of external threats, conditions that have triggered the rise of increasingly defensive modes of homeownership (pp. 26–44).

A key driver of this shift lies in Zuboff's model of *Surveillance Capitalism* [37], in which data harvesting and behavioural monetisation are not incidental but central to corporate power. As Zuboff notes, "*Surveillance capitalists discovered that the most predictive behavioural data come from intervening in the state of play to nudge, coax, tune, and herd behaviour toward profitable outcomes*" (p. 295). By tapping into emotional vulnerabilities and fears of intrusion or crime, companies market smart doorbells as an essential feature of a digitally fortified lifestyle. In doing so, they reinforce perceived threats while expanding corporate influence over domestic life.

Whether capturing location data, behavioural patterns, social interactions, or the movements of a passerby and visitors, everyday life within and around the home is absorbed by smart doorbells and transformed into commodified surveillance inputs. This commodification is neither abstract nor hypothetical. It is contractually embedded in corporate policies. For example, Ring's Terms of Service state: "*You [user] hereby grant Ring and its licensees an unlimited, irrevocable, fee-free and royalty-free right to use, distribute, store, copy, modify, and create derivative works for any purpose [...]*" [28, 29]. While the policy does not explicitly confirm the sale of user data, it permits broad usage for advertising, analytics, and other commercial purposes. This contractual logic reflects Zuboff's concern that both users and those being surveilled are monetised and disempowered within a corporate ecosystem that commodifies not only domestic space but also social behaviour.

3.2 Watching Without Watching: The Panoptic Effect at the Doorstep

While Surveillance Capitalism explains the monetisation of data extracted from smart doorbell surveillance, the psychological and behavioural effects of living under constant observation are best understood through Foucault's concept of *Panopticism* [17].

In this context, smart doorbells extend the logic of the panoptic gaze beyond institutional spaces, embedding surveillance as an ordinary feature of domestic life. Foucault's concept of *Panopticism* [17] describes a surveillance model in which individuals alter their behaviour under the mere assumption of being watched, regardless of whether surveillance is actively occurring [17, 23, 25]. Smart doorbells replicate this dynamic at

the domestic threshold, creating a panoptic effect, where digital overseers continuously monitor and capture any movement, be it gait, facial expressions, conversations, tone of voice, and even habitual walking routes. As one participant from our pilot survey noted, “*Looking into surveillance through smart doorbells made me realise that this changes my very nature when I am out.*” So, the possibility of constant observation compels both residents and passersby to self-regulate their actions, suppress emotional expression, adjust daily routines, and restructure social interactions in response to perceived scrutiny.

The next section details the research methodology, starting with a pilot survey and extending into ongoing data collection through interviews and focus groups.

4 Methodology

This study examines stakeholder perspectives on the ethical and social implications of integrating smart doorbells into residential security. It focuses on the trade-offs between safety, convenience, and privacy erosion, as well as broader impacts such as the expansion of surveillance and its influence on community dynamics and social trust. Rather than aiming for statistical generalisation, the research seeks to understand the diverse ways individuals and communities interact with smart doorbell technologies.

4.1 Research Design and Philosophical Approach

To achieve this aim, the study employs a two-phase qualitative research design, first conducting a pilot survey (*Phase 1*) and then expanding into semi-structured interviews and focus groups (*Phase 2*). Building on this foundation, the research is rooted in an interpretivist paradigm and guided by the inductive approach, allowing themes and patterns to emerge organically rather than being constrained by predefined theoretical constructs. Through this perspective, the interpretivist stance positions participants’ lived experiences, perceptions, and social contexts as essential for understanding the ethical and societal implications of smart surveillance.

4.2 Participants and Sampling Strategies

Given the sensitive nature of surveillance and privacy, ethical protocols have been implemented throughout the study. Ethical Approval was granted by De Montfort University’s Research Ethics Committee prior to data collection.

Phase 1 began with a pilot survey conducted in a Leicester (UK) neighbourhood, where 30 out of 48 households installed smart doorbells. Using purposive sampling, the study engaged participants with direct experience of using these devices. The final dataset comprised responses from 19 households, forming the primary user group. Additionally, purposive sampling included non-primary users, consisting of 5 households without smart doorbells installed and 22 individuals who interacted with smart doorbells indirectly, such as neighbours, passersby, and visitors.

Phase 2 builds on these initial findings by extending the participant pool to approximately 50 to 60 participants drawn from five stakeholder categories: *primary users*

(smart doorbell owners and residents who actively use smart doorbells), *non-primary users* (neighbours, passersby, or incidental subjects affected by the technology), *law enforcement representatives* (including police forces and security personnel), *privacy advocates*, and *technology experts* (including academic researchers). Sampling in Phase 2 combines purposive and snowball strategies to reach representation from both the pilot survey participant categories and hard-to-reach groups, including privacy specialists and law enforcement personnel.

Participant recruitment for Phase 2 is further supported by the Cumbria Neighbourhood Watch Scheme Association, the partner organisation, which has previously partnered with police to distribute smart doorbells to residents affected by domestic violence [21].

4.3 Data Collection Stages

In Phase 1, participants were recruited through door-to-door outreach and local community channels. This pilot survey focused on users' perceptions of smart doorbell use in relation to surveillance, the social fabric of the neighbourhood, and definitions of "suspicious behaviour." Additionally, a supplementary survey for non-primary users captured broader perceptions from those indirectly exposed to these devices. Drawing on these insights, the pilot findings were then used to refine the research question, design instruments for Phase 2, and identify the need for broader and more diverse categories of stakeholders.

Phase 2, currently underway, involves semi-structured interviews and focus groups. These methods were specifically selected to allow for both individual reflection and collective dialogue, facilitating a deeper examination of lived experiences and normative perspectives on smart doorbell surveillance practices.

4.4 Data Analysis

Data from both phases undergo thematic analysis to identify key patterns, contradictions, and emerging ethical concerns across stakeholder groups. For example, in the pilot phase, preliminary themes, including surveillance rivalry, moral ambiguity, and the erosion of community trust, helped refine the design and analytical framework of Phase 2 and shaped interview and focus group protocols to probe deeper into stakeholder tensions and concerns.

5 Pilot Study Findings

This chapter presents the preliminary findings from the pilot survey, organised into themes that emerged from participant responses. The sample included 19 households with smart doorbells, 5 households without them, and 22 non-primary users. Despite its limited size, the dataset offers early insight into the social, ethical, and relational dynamics surrounding smart doorbell adoption. The findings suggest that these devices are beginning to reshape neighbourhood interactions in complex and, at times, troubling ways.

Theme 1. Eroding Trust and Changing Community Dynamics. Participants described how the visible presence of smart doorbells introduced a sense of mutual suspicion among neighbours, particularly when residents became aware that their movements might be recorded by someone else's doorbell. This shift prompted some to withdraw from casual interactions or avoid areas they previously considered neutral communal spaces. What was originally intended to enhance security measures instead undermined informal social trust.

Theme 2. Surveillance in Neighbour Disputes. Beyond security, in some cases, smart doorbells were used as an instrument of surveillance to deter, document, or confront those involved in neighbourhood tensions. For some, these devices became tools to gather evidence in ongoing disputes and, in some cases, tools for asserting dominance and control in strained relationships.

Theme 3. Safety and Protection in Vulnerable Contexts. For a smaller group, smart doorbells provided a sense of protection, particularly in cases of domestic abuse or personal safety concerns. These users relied on footage for legal documentation, reinforcing the dual nature of smart doorbells as both protective mechanisms in high-risk situations and contributors to mistrust elsewhere.

Theme 4. Incidental Recording and Ethical Uncertainty. Participants admitted to saving footage as a precaution or reviewing recordings out of curiosity, even when the content was not clearly relevant to safety and intended for surveillance. Some stated they would share footage with law enforcement upon request, without seeking consent from those captured in the recordings. This underscores the unclear legal and ethical status of such footage. Additionally, the findings highlight a lack of awareness regarding data retention practices, reinforcing the need for greater transparency and user education on surveillance ethics.

Theme 5. Ambivalence Toward Surveillance Normalisation. Responses from non-primary users ranged from indifference to strong opposition to being recorded by video doorbells. Concerns included threats to personal privacy, behavioural self-censorship, surveillance creep, and questions about the legitimacy of private surveilling public spaces.

While some participants saw value in using smart doorbells to deter crime or assist law enforcement by sharing footage, others voiced concerns about the normalisation of surveillance in residential areas. Several questioned whether the balance between security and privacy had been adequately considered in the design and deployment of these technologies and feared that such widespread surveillance might gradually reshape the perception of socially acceptable behaviour.

The next chapter reflects on the key emerging themes through both theoretical and normative lenses.

6 Discussion

The pilot findings reveal that smart doorbells are not only shaping how residents protect their homes, but also transforming how they interact with others, navigate privacy boundaries, and perceive surveillance in everyday life. Participants' narratives highlighted emerging ethical tensions, including shifts in behavioural norms, erosion of

trust, heightened social suspicion, the rise of surveillance-related disputes, and ongoing uncertainty around consent and over privacy norms. These findings suggest that smart doorbells are more than passive security tools; they actively participate in reconfiguring social relations and the ethical contours of neighbourhood life.

6.1 Surveillance and Social Self-regulation

One of the most immediate impacts observed in the pilot study was how smart doorbells influence behavioural self-regulation. Participants reported modifying how they moved, spoke, and interacted when near homes with smart doorbells, such as avoiding conversations, lowering their voices, or limiting contact with others. These accounts suggest a shift in behaviour driven by awareness of being recorded.

Foucault's concept of *Panopticism* [17] provides a lens to understand this transformation. Originating in his analysis of the panopticon, where individuals regulate their behaviour under the assumption of constant observation, the concept extends to modern residential environments where surveillance becomes internalised. Smart doorbells replicate this logic on a microscale, shaping behaviour through the perceived possibility of being watched. Unlike centralised, institutional forms of surveillance, smart doorbell surveillance is decentralised and ambient, distributing the power of observation across neighbourhoods. As a result, individuals in these neighbourhoods exist in a dual role as both *watchers* and *the watched*. This dynamic erodes informal trust, replacing spontaneous interaction with caution and self-conscious navigation of public and semi-private spaces.

6.2 Data Commodification and Platform Power

Zuboff's theory of *Surveillance Capitalism* [37] helps shed light on the broader ecosystem in which smart doorbell surveillance unfolds. While most users purchase these devices primarily for home security, many remain unaware that their recorded footage and behavioural patterns, such as daily routines, activities, and interactions around the property, may be analysed or redistributed by platform providers like Amazon (*Ring*) or Google (*Nest*).

From Zuboff's perspective, smart doorbells do more than secure property entrances, they also turn domestic environments into nodes for behavioural data streams. Moreover, these devices often operate under opaque terms of service that are difficult to navigate and rarely scrutinised by users, leaving them with limited control over how their data is used. Meanwhile, companies retain broad rights to use and repurpose this information. What begins as a personal security device becomes gradually integrated into a wider surveillance infrastructure that commodifies everyday life. This transformation remains largely invisible to the users while serving corporate interests.

6.3 Normative Frameworks: Ethical Design and Governance

Ethical ambiguity emerged as another central theme in the pilot findings. Participants described instances where smart doorbell footage was used in neighbour disputes and

personal conflicts, as well as situations in which bystanders were inadvertently recorded without their knowledge or consent. These scenarios underscore the complex challenges of navigating consent, privacy rights, data ownership, and appropriate use. More broadly, they reflect the normalisation of surveillance, which increasingly leaves users to navigate an uncertain ethical terrain alone.

Ethics by Design offers a pathway for embedding ethical principles into the architecture of smart surveillance technologies. In the absence of such embedded ethics, the burden of managing moral and legal risks falls disproportionately on individuals, many of whom may lack the resources or knowledge to assess these complexities. Rather than leaving ethical decision-making solely to users, this approach encourages the inclusion of features such as real-time alerts, consent requests, and clear usage policies, helping users manage the unintended consequences of domestic surveillance.

Privacy by Design. The pilot also revealed considerable gaps in user awareness and understanding regarding privacy rights of those being audio- and video-recorded by smart doorbells, as well as third-party data access and long-term storage practices. Such oversights point to a broader failure to incorporate robust Privacy by Design principles. Ideally, principles such as privacy protection, data minimisation, transparency in data flows, and greater user control over information should be integrated into smart doorbell functionalities as a core design priority, not an afterthought, in a context where surveillance intersects with private and public life.

Responsible Research and Innovation (RRI). Applying *RRI* to smart doorbell ecosystems means co-developing ethical guidelines, setting surveillance boundaries, and assessing long-term societal impacts before harms materialise. This model promotes a broader governance framework that extends responsibility beyond individual users to include policymakers, designers, and platform providers. It further advocates for anticipation, inclusiveness, and a reflexive approach to technology development.

Designing smart doorbells responsibly requires more than technical efficiency. It demands the integration of ethical principles that anticipate misuse, promote transparency, and protect the rights of both those surveilling and those being surveilled.

7 Conclusion

This research provides a timely contribution to the underexplored domain of smart doorbell surveillance and highlights its effects on privacy, trust, and community dynamics. It contributes to ongoing debates on responsible technology use by examining how individuals and communities negotiate the trade-offs between security, autonomy, and social cohesion in surveillance-driven environments.

The study's contribution is two-fold. Firstly, it presents empirical insights from a UK-based pilot study, shedding light on the lived experiences of smart doorbell users and bystanders. Findings reveal complex and sometimes contradictory effects, including ethical ambiguity in video doorbell footage use, alongside the trade-offs between enhanced security and the growing social fragmentation. These experiences point to a broader normalisation of surveillance at the residential threshold and a shift toward behavioural self-regulation among both device owners and incidental subjects. In several cases, this normalisation led to a "surveillance arms race," where individuals felt

compelled to adopt these devices not just for security but to align with community expectations.

Secondly, the study connects empirical findings to theoretical frameworks of Panopticism and Surveillance Capitalism, alongside normative approaches such as Privacy by Design, Ethics by Design, and RRI. Drawing on Foucault's concept of Panopticism [17], the study demonstrated how the persistent potential for smart doorbell surveillance encourages behavioural discipline and self-regulation, even in the absence of direct observation. In parallel, Zuboff's Surveillance Capitalism exposes how routine doorstep interactions are transformed into commodifiable data assets, often without the awareness or consent of those recorded, thus making ordinary spaces part of a broader corporate surveillance ecosystem.

The discussion further proposes actionable responses, advocating for the integration of Privacy by Design, Ethics by Design, and RRI into the design and governance of smart doorbell technologies. Embedding these normative frameworks would mitigate privacy risks while promoting transparency and responsible technology use, ensuring better protection for both users and those subject to their gaze.

7.1 Limitations and Future Work

This paper represents the first phase of a two-stage research project. While the pilot survey provides valuable initial insights, the relatively small sample size of 19 households limits the generalisability of these findings. The forthcoming qualitative phase, with a more extensive and diverse participant pool, including privacy specialists, law enforcement, and technology experts, is designed to address these limitations. This future work aims to deepen and diversify the analysis of smart doorbell surveillance and validate the preliminary themes, thereby further informing the more effective policies and design-oriented strategies for responsible technology use.

Acknowledgements. This research is funded by the UK Research and Innovation (UKRI).

Engineering and Physical Sciences Research Council (EPSRC) under Grant No. EP/S023305/1.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alshehri, A., Spielman, J., Prasad, A., Yue, C.: Exploring the privacy concerns of bystanders in smart homes from the perspectives of both owners and bystanders. *Proc. Priv. Enhancing Technol.* **2022**(3), 99–119 (2022)
2. Atkinson, R.: *Domestic Fortress: Fear and the New Home Front*, pp. 26–44, 107–131, 175–193. Manchester University Press, Manchester (2016)
3. BBC News: Plea for doorbell footage after tree attacks, BBC News, 10 February. <https://www.bbc.co.uk/news/articles/c5y7jlj9x0po>. Accessed 10 Feb 2025

4. Beer, D.: The Age of Surveillance Capitalism: The curious philosophy of Amazon's video doorbell shows the tech giants have already won the debate, *Prospect Magazine*. <https://www.prospectmagazine.co.uk/ideas/technology/40093/the-curious-philosophy-ofamazons-video-doorbell-shows-the-tech-giants-have-already-won-the-debate>. Accessed 5 Jan 2025
5. Berndtsson, M., Hansson, J., Olsson, B., Lundell, B.: *Thesis Projects: A Guide for Students in Computer Science and Information Systems*, 2nd edn. Springer, London (2008)
6. Bouchrika, I.: A survey of using biometrics for smart visual surveillance: Gait recognition, in *Surveillance in Action: Technologies for Civilian, Military and Cyber Surveillance*, pp. 3–23 (2018)
7. Bridges, L.: Amazon's Ring is the largest civilian surveillance network the US has ever seen. *The Guardian*, 18 May. <https://www.theguardian.com/commentisfree/2021/may/18/amazon-ring-largest-civiliansurveillance-network-us>. Accessed 25 Nov 2024
8. Bryman, A.: *Social Research Methods*, 4th edn. Oxford University Press, Oxford (2012)
9. Buil-Gil, D., et al.: The digital harms of smart home devices: a systematic literature review. *Comput. Hum. Behav.* **145**, 107770 (2023). <https://doi.org/10.1016/j.chb.2023.107770>. Accessed 23 Feb 2025
10. Burkholder, G.J., Cox, K.A., Crawford, L.M., Hitchcock, J.H.: *Research Design and Methods: An Applied Guide for the Scholar Practitioner*. SAGE Publications, Thousand Oaks (2019)
11. Calacci, D., Shen, J.J., Pentland, A.: The Cop in your neighbor's doorbell. In: *Proceedings of the ACM on Human-Computer Interaction*, pp. 1–19 (2022). <https://doi.org/10.1145/3555125>. Accessed 28 Jan 2025
12. Dawson, C.: *Practical Research Methods*. How To Books Ltd., Oxford (2002)
13. Denscombe, M.: *The Good Research Guide: Research Methods for Small-Scale Social Research Projects*. 7th edn. McGraw-Hill Education, Maidenhead (2021)
14. Doctorow, C.: *How to Destroy Surveillance Capitalism: Medium Editions*, Medium Editions (2021)
15. Fairhurst v. Woodard. Judgment of the Oxford County Court (2022). <https://www.judiciary.uk/wp-content/uploads/2022/07/Fairhurst-v-Woodard-Judgment-1.pdf>. Accessed 12 Nov 2024
16. Fantozzi, C., Grigoletto, L.: At Home With Us. In: *6th EAI International Conference on Smart Objects and Technologies for Social Good*, pp. 1–12. Springer, Heidelberg (2020). <https://doi.org/10.1145/3411170.3411273>. Accessed 31 Dec 2024
17. Foucault, M.: *Discipline and Punish: The Birth of the Prison*, new Penguin, London (1991)
18. Grauer, Y.: Video Doorbell Cameras Record Audio, Too. *Consumer Reports*. <https://www.consumerreports.org/video-doorbells/video-doorbell-cameras-record-audio-too/a4636115889/>. Accessed 2 Apr 2025
19. Kelly, K.A.: The ring video doorbell and the entry of amazon into the smart home: implications for consumer-initiated surveillance. *J. Consum. Policy* **46**, 95–104 (2023). <https://doi.org/10.1007/s10603-022-09534-3>. Accessed 10 Dec 2024
20. Kurwa, R.: Building the digitally gated community: the case of nextdoor. *Surveill. Soc.* **17**(1/2), 111–117 (2019). <https://doi.org/10.24908/ss.v17i1/2.12927>
21. Maalsen, S., Sadowski, J.: The smart home on FIRE: amplifying and accelerating domestic surveillance. *Surveill. Soc.* **17**, 118–124 (2019). Accessed 24 Jan 2025
22. Magnusson, E., Marecek, J.: *Doing Interview-based Qualitative Research: A Learner's Guide*. Cambridge University Press, Cambridge (2015). <https://doi.org/10.1017/9781107449893>. Accessed 13 Nov 2024
23. Manokha, I.: Surveillance, panopticism, and self-discipline in the digital age. *Surveill. Soc.* **16**(2), 219–237 (2018). Accessed 12 Feb 2025
24. Morris, J.: Surveillance by amazon: the warrant requirement, tech exceptionalism, & ring security. *BUJ Sci. Tech. L.* **27**, 237 (2021). Accessed 16 Mar 2025

25. Neighbourhood Watch: Successful Video Doorbell Project. <https://www.ourwatch.org.uk/successful-video-doorbell-project> (n.d.) Accessed 23 May 2024
26. Oates, B.J.: *Researching Information Systems and Computing*. SAGE Publications, London (2006)
27. Reid, L., Sisel, G.: Digital care at home: exploring the role of smart consumer devices. *Health Place* **73**, 102961 (2022). <https://doi.org/10.1016/j.healthplace.2022.102961>. Accessed 30 Nov 2024
28. Ring LLC: Ring Terms of Service (2025). <https://en-uk.ring.com/pages/terms>. Accessed 28 May 2025
29. Ring: Privacy Notice (2025). <https://en-uk.ring.com/pages/privacy-notice>. Accessed 28 May 2025
30. Sadowski, N.: *Too Smart: How Digital Capitalism Is Extracting Data, Controlling Our Lives, and Taking over the World*. MIT Press, New York (2020). <https://doi.org/10.7551/mitpress/12240.001.0001>
31. Salaudeen, M.: Panopticism and Surveillance Capitalism, Medium. <https://medium.com/@mustapher07/panopticism-and-surveillance-capitalism4d6701519429>. Accessed 28 Dec 2024
32. Schleusener, S.: The Surveillance Nexus: digital culture and the society of control. *REAL: Yearbook Res. English Am. Lit.* **34**(1), 175–202 (2019). Accessed 22 Jan 2025
33. Selinger, E., Durant, D.: Amazon's ring: surveillance as a slippery slope service. *Sci. Culture* **31**(4), 1–15 (2021). <https://doi.org/10.1080/09505431.2021.1983797>. Accessed 18 Nov 2024
34. Tabassum, M., Kropczynski, J., Wisniewski, P., Richter Lipford, H.: Smart home beyond the home: a case for community-based access control. In: 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12. ACM, New York (2020) <https://doi.org/10.1145/3313831.3376255>. Accessed 25 Nov 2024
35. Townsend, M.: How CCTV played a vital role in tracking Sarah Everard – and her killer. *The Guardian*, 2 October. <https://www.theguardian.com/uknews/2021/oct/02/how-cctv-played-a-vital-role-in-tracking-sarah-everard-and-her-killer>. Accessed 13 Dec 2024
36. Wong, R.Y., Valdez, J.C., Alexander, A., Chiang, A., Quesada, O., Pierce, J.: Broadening privacy and surveillance: eliciting interconnected values with a scenarios workbook on smart home cameras. In: 2023 ACM Designing Interactive Systems Conference (DIS '23), pp. 1–12. ACM, New York (2023) <https://doi.org/10.1145/3563657.3596012>. Accessed 15 Dec 2024
37. Zuboff, S.: *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York (2019)



User Engagement and Barriers in the Standardization Processes of Digital ID Architectures

Yoshiaki Fukami¹(✉)  and Kazue Sako² 

¹ Tokyo University of Science, 1-11-2 Fujimi, Chiyoda-ku, Tokyo 102-0071, Japan
yoshiaki@rs.tus.ac.jp

² Waseda University, 3-4-1 Ookubo, Shinjuku-ku, Tokyo 169-8555, Japan
kazuesako@aoni.waseda.jp

Abstract. Digital identity systems are increasingly central to both public and private services, yet the standardisation processes underpinning them remain fragmented. This paper examines the ethical, institutional, and governance challenges that arise from the distributed nature of digital identity specification development. Drawing on an interpretive case study approach, we analyse five key communities—namely IETF, W3C, OpenID Foundation, the Decentralized Identity Foundation, and the Internet Identity Workshop—to understand how technical norms are created and how engineers and public actors engage with them. Our findings show that while formal standardisation bodies provide structured processes, public institutions face significant barriers to participation, including technical complexity, fragmented venues, and limited institutional coordination. Informal forums like Internet Identity Workshop (IIW), though inclusive in ethos, pose accessibility challenges due to their high-context structure. We argue that effective governance of digital identity standards requires new forms of coordination, interdisciplinary collaboration, and shared institutional capacity across technical and policy domains. Rather than advocating for centralised administration, we call for constructive interoperability between stakeholder communities to ensure that evolving standards align with public values and democratic oversight.

Keywords: Digital Identity · ID architecture · standardization · data governance · multi-stakeholder

1 Introduction

Digital identity systems (DIS) are increasingly central to how individuals access public and private services. These systems aim to enhance security, streamline service delivery, and protect privacy. However, the standardisation processes underpinning DIS remain fragmented, with limited coordination among technologists, policymakers, and legal experts [1].

Traditionally, standardisation followed a linear path: technical specifications were developed by expert communities, formally recognised by de jure bodies such as the

International Organization for Standardization (ISO) or the International Telecommunication Union Telecommunication Standardization Sector (ITU-T), and only then adopted by public authorities under regulatory frameworks. This sequential model allowed for stable governance and regulatory oversight prior to deployment.

Today, this logic is increasingly disrupted. Innovations such as decentralised identifiers, cryptographic credentials, and identity wallets evolve rapidly within open technical communities—often preceding legal codification or formal institutional adoption. Standardisation by consortia such as the World Wide Web Consortium (W3C), the Decentralized Identity Foundation (DIF), and OpenID Foundation now unfolds in parallel with public-sector implementation, blurring the lines between specification and policy.

Many specifications with broad societal implications are now shaped in decentralised venues lacking formal review mechanisms or institutional anchoring. Fragmentation across jurisdictions exacerbates coordination difficulties, hindering the development of coherent, user-centric identity frameworks. Regulatory tools like sandboxes have been introduced to manage such uncertainty in areas including smart city technologies [2]. These frameworks allow policymakers to test new technologies in controlled environments while addressing privacy and ethical concerns.

Various ethical guidelines have been introduced to govern the responsible development of digital technologies. The UNESCO Recommendation on the Ethics of Artificial Intelligence emphasizes human-centered design and responsible innovation [3]. The OECD AI Principles highlight transparency, accountability, and the protection of human rights [4]. The IEEE's Ethically Aligned Design outlines the need to prioritize human well-being in the development of autonomous systems [5].

However, one of the key challenges in the governance of emerging digital technologies is addressing the responsibility gap that arises in the deployment of autonomous systems. Gotterbarn and Wolf [6] explore this issue by examining how the ACM Code of Ethics can provide a framework for clarifying accountability in AI systems. They argue that the responsibility gap—where it is unclear who should be held accountable for the actions of autonomous systems—poses significant ethical risks.

Digital identity systems by governments worldwide aim to improve service delivery, ensure security, and enhance digital transformation. For example, the EU's Digital Identity Wallet (EUDIW) provides citizens with a secure platform for managing cross-border services while adhering to the General Data Protection Regulation (GDPR) [7]. Similarly, Singapore's SingPass system offers integrated access to over 2,000 services across public and private sectors, incorporating biometric authentication to enhance security [4].

Digital identity is a particularly salient case of this evolving model. Governments are increasingly deploying systems—such as EUDIW or Japan's MyNumber—while relevant standards are still under active development. Public authorities are no longer passive adopters but participants in ongoing technical processes, often without clear institutional mandates or engagement strategies.

This paper examines the ethical, institutional, and governance implications of this non-linear standardisation model. It argues for improved mechanisms of coordination and oversight to ensure that evolving technical norms align with public values, legal protections, and inclusive participation.

2 Background and Related Work

2.1 Digital Identity Systems and Governance Challenges

Digital identity refers to the collection of attributes, identifiers, and credentials that represent an entity—typically a person—in digital environments. These systems form the backbone of trust and access in digital transactions and are increasingly essential for both public and private service delivery.

Earlier digital identity models relied on service-specific credentials, such as usernames and passwords. While easy to deploy, these mechanisms have long suffered from poor usability and weak security, leading to risks such as password reuse, fragmentation, and impersonation attacks [8, 9]. Federated identity systems—such as those using OpenID Connect or OAuth 2.0—were introduced to address these limitations by enabling single sign-on through trusted third-party providers. However, they introduced new concerns related to centralisation and platform dependency.

To mitigate such issues, decentralised identity architectures have emerged. Standards such as Decentralized Identifiers (DIDs) and Verifiable Credentials (VCs)—developed by communities including the W3C, DIF, and the OpenID Foundation—seek to enable interoperable, privacy-preserving credential exchange with more user control. These efforts support cross-domain credentialing without relying on centralised authorities, offering new pathways for identity portability.

Yet technical innovation alone cannot ensure responsible adoption. The deployment of decentralised identity systems requires alignment across technical, legal, and institutional domains. This includes cryptographic implementation, data protection regulation (e.g. GDPR), procurement policy, and trust frameworks for issuers and verifiers. Interoperability, legal compliance, and institutional accountability must be jointly considered at early stages of design and adoption.

Some stakeholders argue that using blockchain for digital identity systems poses risks related to privacy, scalability, and environmental sustainability [10]. For instance, the immutability of blockchain records can conflict with the right to be forgotten under GDPR [11]. This issue raises questions about how personal data should be managed in decentralized systems, particularly in light of privacy regulations [12].

A critical challenge is the shortage of interdisciplinary expertise. Technical communities and public-sector actors often operate in parallel, with limited overlap in knowledge, resources, and communication practices. This gap can result in systems that are technically robust but institutionally fragile—unable to satisfy ethical, legal, or operational requirements when scaled in public infrastructure.

2.2 Standardization Models in the Digital Domain

The development of technical standards for digital identity does not follow a single, linear process. Rather, it unfolds through a dynamic and distributed ecosystem involving engineering communities, academic researchers, and industry-led consortia. In contemporary ICT domains—particularly the internet and the web—standards often emerge from *de facto* processes in which draft specifications are openly discussed and tested

before being formally adopted by bodies such as the International Organization for Standardization (ISO) or the International Telecommunication Union (ITU-T).

A variety of actors contribute to this model. Organizations such as the Internet Engineering Task Force (IETF), the World Wide Web Consortium (W3C), the OpenID Foundation (OIDF), and the Decentralized Identity Foundation (DIF) have each played roles in shaping technical specifications relevant to digital identity, including data models, credential formats, and authentication mechanisms. These groups operate with varying levels of formalization and openness, ranging from individual-based communities (e.g., IETF) to membership-driven consortia (e.g., W3C, OIDF).

While none of these bodies possess regulatory authority, their outputs often serve as the technical foundations for government-led identity infrastructure. Notably, mechanisms such as ISO's Fast Track Procedure allow specifications developed in these venues to be formally recognized as international standards, although many widely used protocols never go through this process. This demonstrates that formal ratification is not a prerequisite for broad adoption.

In recent years, the acceleration of innovation and the rise of digital public infrastructure have changed how governments interact with standardization. Public-sector actors now engage not only in adopting established standards, but also in shaping early specifications and piloting implementations. This shift blurs the line between development and deployment, demanding new institutional capacities for monitoring and technical engagement.

In decentralized identity, standard governance is particularly complex. Stakeholders must navigate overlapping specifications, differing trust models, and varied implementation practices. Features like selective disclosure or credential revocation may be specified but inconsistently realised, raising concerns about interoperability, legal compliance, and privacy.

Digital identity standardization must therefore be seen as a socio-technical process—one shaped by negotiation between developers, governments, and global institutions. Ensuring meaningful participation requires not only technical skill, but also mechanisms for inclusion, accountability, and responsiveness to societal values.

2.3 Ethical and Institutional Considerations in Decentralized Identity Standardization

As decentralized identity systems become more integrated into public and private infrastructure, ethical and institutional challenges arise alongside technical ones. These challenges concern not only implementation, but the very processes through which standards are developed and adopted.

Despite principles of openness, venues like W3C and DIF present practical barriers: high technical demands, rapid iteration, and scheduling or language constraints. While mailing lists and specifications are public, meaningful participation often requires domain expertise and institutional support, limiting access for civil society and smaller public-sector actors.

Technical decisions—such as support for unlinkability or selective disclosure—shape how user rights like privacy are preserved. Mechanisms such as zero-knowledge

proofs exist, but adoption is uneven. For instance, the EU Digital Identity Wallet uses linkable identifiers for auditability, potentially compromising privacy.

Decentralization distributes control, but can obscure accountability unless roles for issuers, holders, and verifiers are well-defined. Most standard bodies, like IETF and W3C, do not define governance models for deployment. While many host privacy-focused groups or ethics liaisons, these functions support but do not replace legal frameworks. Public authorities thus face the challenge of translating technical artefacts into enforceable policies, often under time and capacity constraints.

Inclusion remains uneven, not due to a lack of concern, but due to gaps in operational integration. Ethical design, institutional accountability, and broader participation must be embedded more deliberately in the standardization lifecycle—especially as these technologies underpin essential public services.

2.4 Ethical Considerations and Governance Implications

As digital identity systems move toward more decentralised models, ethical and governance issues increasingly intersect with technical design. These infrastructures—comprising issuers, holders, verifiers, and intermediaries—require not only interoperability but also clearly defined responsibilities and protections.

A key challenge lies in how governance assumptions are encoded into technical specifications. While features like selective disclosure and decentralised revocation support privacy by design, their effectiveness depends on implementation by issuers and relying parties. Without coordination, implementations may diverge from normative goals.

Standardisation bodies such as W3C and IETF primarily focus on technical interoperability. Yet they also host internal groups—like W3C’s Privacy Interest Group or the IAB’s Ethics and Policy Group—to surface ethical concerns and promote value-sensitive design. These groups carry weight, but their remit is limited: they do not define regulatory frameworks or resolve legal trade-offs.

This limitation becomes critical when public-sector agencies adopt emerging standards in high-stakes domains. Design decisions made early in technical communities can shape later issues of accountability, data governance, and rights. Without mechanisms for interpretation and institutional anchoring, decentralised identity systems risk drifting from public values. Rather than imposing centralised oversight, what is needed are coordination frameworks that align technical design with ethical governance—through better interpretability, cross-domain collaboration, and support for institutional capacity in both engineering and regulatory communities.

3 Research Design and Analytical Scope

This paper adopts an interpretive case study approach [13] to examine how digital identity standards are shaped by technical communities. Rather than aiming for generalisability or hypothesis testing, this study focuses on building a situated understanding of institutional configurations and stakeholder interactions as they unfold in practice. As argued by Welch et al. [14], the interpretive case method is particularly suited for analysing dynamic, context-specific developments that involve multiple perspectives and high levels of uncertainty.

Given the accelerating pace of technological convergence—driven by developments in AI, the Internet of Things, and government-led data policies—digital identity infrastructure has become both more fragmented and more interdependent. The increasing ratio of software components in infrastructure, combined with lower implementation costs due to internet-based interoperability testing, has led to a broader trend of overlapping standardisation and deployment phases. However, these structural shifts have gained particular momentum in recent years, requiring careful attention to the emerging coordination practices among public, private, and civil actors.

To understand these processes, the analysis draws on publicly available governance documents, participation in community meetings and calls, and secondary materials including websites and social media posts.

The development of digital identity systems is shaped by a constellation of technical communities and collaborative venues. This section examines four standardisation bodies—IETF, W3C, OIIF, and DIF—which formally develop specifications through structured working groups. In contrast, we also discuss the Internet Identity Workshop (IIW), a community forum that supports informal dialogue and early-stage prototyping but does not produce specifications. Understanding their respective roles is critical for analysing how standards emerge, evolve, and influence digital identity infrastructures.

4 Case Study: Technical Communities in Digital Identity Standardisation

4.1 Formal Specification Bodies: IETF, W3C, OIIF, and DIF

These four organisations serve as the principal venues for the development of digital identity-related technical specifications. Each operates working groups that follow defined procedures to produce, review, and publish specifications (Table 1).

These organisations differ in governance and openness, as shown in Table 1. These venues not only support formal specification development but also coordinate with each other. For example, OIIF's OpenID Connect builds on OAuth 2.0 defined by IETF, while DIF's preliminary work informs W3C's formalisation of VC-related standards. In this sense, their activities are interlinked, forming an iterative pipeline from prototyping to technical formalisation and, in some cases, to regulatory consideration.

These organisations differ in governance and openness, as shown in Table 1. These venues not only support formal specification development but also coordinate with each other. For example, OIIF's OpenID Connect builds on OAuth 2.0 defined by IETF, while DIF's preliminary work informs W3C's formalisation of VC-related standards. In this sense, their activities are interlinked, forming an iterative pipeline from prototyping to technical formalisation and, in some cases, to regulatory consideration.

4.2 Informal Dialogue and Early Innovation: The Role of IIW

The Internet Identity Workshop (IIW) differs fundamentally from the above institutions. It is not a standard-setting organisation, but an unconference-style event held semi-annually in California. IIW brings together developers, policymakers, civil society, and industry representatives to engage in open discussion, demo sessions, and peer learning.

Table 1. Differences in governance and openness among organizations

Organisation	Core Focus	Specification Output	Participation Model	Key Identity Standards
IETF	Internet protocols	Individual-based, open	RFCs	OAuth 2.0, JWT
W3C	Web standards	Institutional membership + invited experts	Recommendations	DID, VC, WebAuthn
OIDF	Identity protocols	Member-based, consensus-driven	Versioned specs	OpenID Connect, FAPI
DIF	Decentralized identity tooling	Open with Contributor Agreement	Drafts, libraries	DIDComm, Universal Resolver

While IIW does not operate working groups or publish specifications, it plays a vital incubatory role in the standardisation ecosystem. It provides a space for cross-sector dialogue, early alignment on emerging concepts, and interoperability testing of experimental implementations. Technologies such as DIDs, Verifiable Credentials, and wallet prototypes were discussed and prototyped at IIW before formal specification in other venues.

However, IIW’s informal structure and high-context participation model can pose barriers to broader engagement. The event’s physical location, use of English, and reliance on community familiarity contribute to geographic and institutional asymmetries. Nonetheless, its openness and spontaneity serve as a critical complement to the procedural formality of other standardisation forums.

4.3 Implications for Governance and Coordination

Understanding the distinctions between these actors is essential for public institutions seeking to engage with technical standardisation. Formal specification bodies like IETF, W3C, OIDF, and DIF offer stable procedures for standard development and institutional accountability. In contrast, IIW offers early insights and normative framing but lacks continuity mechanisms or ratification processes.

Rather than forming a linear hierarchy, these entities constitute a distributed and layered standardisation ecosystem. Effective public-sector participation thus requires the capacity to navigate both formal and informal venues, assess the provenance of technical artefacts, and engage across multiple layers of the specification lifecycle.

5 Discussion

5.1 Rethinking Standardisation in Digital ID: From De Jure Authority to Distributed Engagement

The foundational infrastructure of the Internet and the World Wide Web has long been shaped through collaborative processes involving academic institutions, industry actors, and independent technologists. Standard-setting activities have been coordinated through bodies such as IETF and W3C, while actual implementation and operational governance have remained decentralised, executed by a diverse array of public and private sector entities. This decentralised and consensus-driven model has, in many cases, facilitated rapid innovation and promoted interoperability across heterogeneous systems, though the normative consequences of this model are still being debated.

In recent years, governments and public-sector institutions have become increasingly reliant on internet- and web-based infrastructures for the provision of public services, including identity verification, eligibility assessment, and delivery of benefits or administrative authorisations. These functions often depend on digital identity systems that build upon specifications actively discussed or developed within distributed standardisation venues.

As public authorities seek to design, adopt, or regulate such digital identity systems, they face growing pressure to understand and interact with technical specifications whose development processes are fragmented, iterative, and technically demanding.

The key challenge is not simply the decentralisation of standard-setting, but the rising complexity and scope of engagement required from public actors to evaluate, influence, or align with these emerging technical norms—often in real time, and with limited institutional capacity. This paper highlights the ethical and governance implications of this evolving landscape, and underscores the need for new mechanisms that support inclusive, transparent, and accountable standardisation practices that can serve both public interest and technical rigour.

5.2 Ethical and Institutional Gaps in Distributed Standardisation

The decentralised and multi-venue character of contemporary digital identity standardisation presents a number of ethical and institutional challenges. Unlike legacy telecommunication standards or safety-critical systems—where formal public oversight was often built into early stages of technical design—many of today's identity specifications originate in communities that prioritise agility, modularity, and implementation orientation. This approach has proven effective in fostering interoperability and innovation [1], but it complicates systematic consideration of legal compliance, institutional responsibility, and broader public values.

A key friction arises from the increasing reliance of public-sector actors on technical artefacts whose design trajectories are governed by distributed and rapidly iterating communities. As discussed in Sect. 4, organisations such as IETF, W3C, OIDF, DIF, and IIW contribute to different aspects of the evolving identity stack. Yet, their openness varies not only in terms of formal participation but also in practical accessibility. For example, IETF and W3C regularly rotate their meetings across global locations and

publish comprehensive archives; DIF offers open calls and GitHub transparency; by contrast, the Internet Identity Workshop (IIW) has been consistently held in California¹, creating geographic and institutional participation asymmetries. While none of these venues is formally closed, the cumulative effect is a complex landscape in which effective participation requires significant time, technical expertise, and often organisational backing.

Moreover, although many of these organisations have formal provisions for public input or ethical engagement—such as W3C’s Privacy Interest Group, DIF’s governance working groups, or OIDF’s public review mechanisms—the critical question is how such mechanisms are operationalised. Having roles or committees on paper is not sufficient if they do not translate into concrete influence over technical trajectories or implementation norms. Rather than a lack of structure, the challenge is often one of integration: how to meaningfully embed ethical and regulatory concerns into a process architecture optimised for technical efficiency.

The issue is further complicated when public-sector agencies experiment with or deploy systems based on early-stage standards. When national digital ID systems begin to incorporate specifications such as Verifiable Credentials or decentralised identifiers before these have stabilised or reached wide deployment, gaps in legal clarity, rights protection, and accountability mechanisms become pronounced. The tension is not simply a matter of speed, but of procedural misalignment between technical prototyping and public-sector governance (Table. 2).

These observations suggest the need for neither full centralisation nor legal intervention into every layer of technical development. Rather, what is called for is institutional interface capacity: mechanisms, norms, and shared languages that can mediate between distributed standardisation and public-sector responsibilities. The ethical concern lies not in who sets the standards, but in whether the conditions for responsible adoption and adaptation are in place—particularly where fundamental rights, welfare delivery, or public accountability are at stake.

Table 2. Friction Points Between Technical and Regulatory Domains

Stage of Specification	Technical Community Norms	Challenges for Public-Sector Integration
Early Specification Design	Iterative drafting, code-first experimentation	Lack of legal/ethical input; unclear rights frameworks
Testing and Prototyping	Developer-led implementation trials	Minimal regulatory visibility or policy dialogue
Adoption and Maintenance	Ongoing versioning and community governance	Lagging public-sector adaptation and compliance tools

¹ Sometimes there is Satellite IIW in other regions of the world, but not on a periodical basis.

5.3 Towards Integrative Governance Models

As the preceding analysis has shown, digital identity standardisation today unfolds across a variety of decentralised venues, where specification work is largely driven by private or community-based actors. However, the growing adoption of digital identity technologies by public-sector institutions—particularly in domains involving authentication, service eligibility, and legal compliance—has introduced new stakes. These developments challenge existing boundaries between technical standardisation and public governance, raising important questions about accountability, transparency, and procedural legitimacy.

While some government-affiliated entities have begun to observe or occasionally participate in standard-setting activities, this involvement remains limited in both scope and continuity. Formal membership in technical organisations such as W3C, OpenID Foundation, or DIF by public agencies remains the exception rather than the norm. Even when participation does occur, it is often tied to specific pilot projects or short-term objectives, rather than forming part of a sustained institutional strategy. As such, it is premature to describe the current situation as a transition toward a co-development regime.

Nevertheless, the increasing policy relevance of these standards invites exploration of new forms of coordination between technical and public actors. Rather than proposing definitive institutional mechanisms, this paper suggests the importance of further inquiry into how governance interfaces might be designed to improve mutual intelligibility and accountability. These interfaces may involve procedural transparency, normative framing at early stages of technical development, or clearer documentation practices that support external scrutiny.

In particular, a central challenge lies not in simplifying technical or legal documentation—an aim that may be unrealistic given the inherent complexity of both domains—but rather in developing pragmatic methods of interdisciplinary collaboration.

Importantly, the goal is not to subsume technical processes under regulatory authority, nor to impose a uniform model of governance across all standard-setting bodies. Rather, the challenge lies in identifying context-appropriate models of interaction that reflect the differentiated mandates, responsibilities, and institutional logics of each stakeholder group.

In summary, this paper highlights the need for reflexive and integrative governance approaches that can bridge the procedural and epistemic gaps between distributed technical communities and public-sector institutions. Future research should investigate in greater depth how such models might be designed, institutionalised, and adapted to the evolving dynamics of digital identity infrastructure.

6 Conclusion and Future Directions

This paper has examined the evolving standardisation landscape for digital identity technologies, with particular attention to how the increasing technical complexity and institutional diversity of the ecosystem affect governance, transparency, and public-sector engagement. Through a comparative analysis of initiatives by IETF, W3C, OpenID Foundation, DIF, and IIW, we have shown that many of today's digital identity-related

specifications are developed within decentralised and multi-stakeholder settings, rather than through centralised processes managed by intergovernmental bodies alone.

These specification venues—ranging from formal standard development organisations to informal, community-led workshops—generally promote open access and transparency, but differ considerably in their levels of procedural rigour, institutional affiliation, and mechanisms for external accountability. While this distributed model of development has supported innovation and interoperability across a broad array of internet and web technologies, it also presents challenges for public-sector actors tasked with adopting and governing these systems. As governments and public institutions increasingly rely on technical artefacts—such as identity wallets, verifiable credential schemas, and cross-domain authentication protocols—to deliver essential services, they must grapple with specification landscapes that are not always easy to interpret or systematically engage with.

Rather than proposing a return to centralised or exclusively *de jure* standardisation, this paper argues for the importance of constructing deliberate, functional interfaces between technical and policy domains. The question is not where standardisation should occur, but how diverse stakeholders can participate meaningfully in processes that affect public interest outcomes. This includes not only fostering procedural openness and documentation clarity, but also developing sustained, role-based coordination across technical and institutional domains.

From an ethical standpoint, the priority is to design governance approaches that respect the autonomy of technical communities while ensuring that systems with societal impact are subject to context-sensitive review, public reasoning, and accountable stewardship. The aim is not uniform control, but rather constructive interoperability—between codebases and institutions, between engineering logics and public values.

Acknowledgment. This work was supported by SECOM Science and Technology Foundation.

References

1. Fukami, Y.: Open and clarified process of compatibility standards for promoting data exchange. *Rev. Socionetwork Strat.* **15**(3), 1–19 (2021). <https://doi.org/10.1007/s12626-021-00087-4>
2. Yarime, M.: Governing data-driven innovation for sustainability: opportunities and challenges of regulatory sandboxes for smart cities. In: Fukami, Y., Kokuryo, J. (eds.) *AI for Social Good* Association of Pacific Rim Universities Limited (APRU) and Keio University, pp. 180–202 (2020)
3. UNESCO.: Recommendation on the Ethics of Artificial Intelligence (2022). <https://unesdoc.unesco.org/ark:/48223/pf0000381137.locale=en> Accessed 13 Jan 2025
4. OECD.: Recommendation of the council on artificial intelligence (2019). <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449> Accessed 13 Jan 2025
5. IEEE Standards Association Ethically Aligned Design: a Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (2017). https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf Accessed 13 Jan 2025

6. Gotterbarn, D., Wolf, M.J.: Addressing the AI responsibility gap with the ACM code of ethics. In: Arias Oliva, M., Pelegrín Borondo, J., Murata, K., Souto Romero, M. (eds.) *The Leading Role of Smart Ethics in the Digital World*, pp. 177–188 (2024)
7. European Commission Ethics guidelines for trustworthy AI (2021). <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> Accessed 13 Jan 2025
8. Florêncio, D., Herley, C.: A large-scale study of web password habits. In: 16th International Conference on World Wide Web (WWW 2007), pp. 657–666. ACM, New York (2007). <https://doi.org/10.1145/1242572.1242661>
9. Bonneau, J., Herley, C., van Oorschot, P.C., Stajano, F.: The quest to replace passwords: a framework for comparative evaluation of web authentication schemes. In: IEEE Symposium on Security and Privacy (SP 2012), pp. 553–567. IEEE, San Francisco (2012). <https://doi.org/10.1109/SP.2012.44>
10. Linn, L. A., Koo, M.B.: Blockchain for health data and its potential use in health IT and healthcare-related research. In: *ONC/NIST use of Blockchain for Healthcare and Research Workshop*, pp. 1–10. Gaithersburg, Maryland, ONC/NIST (2016)
11. Finck, M.: *Blockchain and the General Data Protection Regulation: Can distributed ledgers be squared with European data protection law?* European Parliamentary Research Service (EPRS) (2018). [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634445/EPRS_STU\(2019\)634445_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634445/EPRS_STU(2019)634445_EN.pdf) Accessed 13 Jan 2025
12. Janssen, H., Cobbe, J., Norval, C., Singh, J.: Decentralized data processing: personal data stores and the GDPR. *International Data Privacy Law* **10**(4), 356–384 (2020). <https://doi.org/10.1093/idpl/ipaa016>
13. Stake, R.E.: Qualitative Case Studies. In: Denzin, N.K., Lincoln, Y.S. (eds.) *The Sage Handbook of Qualitative Research*, 3rd edn., pp. 443–466. Sage Publications, Thousand Oaks (2005)
14. Welch, C., Piekkari, R., Plakoyiannaki, E., Paavilainen-Mäntymäki, E.: Theorising from case studies: towards a pluralist future for international business research. In: Eden, L., Nielsen, B.B., Verbeke, A. (eds.) *Research Methods in International Business*. JIBS Special Collections. Palgrave Macmillan, Cham (2020). https://doi.org/10.1007/978-3-030-22113-3_9



Systematic Review on AI Ethics in Privacy for V2X Communication

Héctor Dave Orrillo Ascama^{1,2,3} , Laercio Cruvinel Júnior^{1,2} ,
and Mário Marques da Silva^{1,2,3} 

¹ Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa,
1169-023 Lisboa, Portugal

{horrillo, lcruvinel, mmsilva}@autonoma.pt

² Autónoma TechLab, 1169-023 Lisboa, Portugal

³ Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal

Abstract. The advent of fully autonomous vehicles (AVs) promises to reshape urban mobility improving efficiency and safety. Despite this, the dependency of AVs on Vehicle-to-Everything (V2X) communication networks introduces substantial cybersecurity vulnerabilities, including remote intrusions and data manipulation. This systematic review investigates strategies designed to safeguard data privacy and integrity within V2X communications, emphasizing the incorporation of ethical principles derived from Artificial Intelligence (AI) Ethics. The analysis indicates that while technical approaches, such as blockchain (BL) and Federated Learning (FL) are extensively adopted, the integration of ethical frameworks remains limited. In response, we propose twelve normative ethical principles aligned with international standards (e.g., IEEE, European Commission) to inform the responsible development and governance of AI systems in AVs. Our results demonstrate that the combination of BL and FL presents strong potential for developing secure, decentralized, and privacy-preserving V2X communication infrastructures.

Keywords: AI applications · autonomous vehicles · blockchain · urban mobility

1 Introduction

Autonomous vehicles (AVs) signify a paradigm shift in urban mobility, offering potential improvements in safety, operational efficiency, and user convenience. However, the growing reliance on V2X communication networks introduces significant vulnerabilities, particularly in relation to data privacy and security. These risks necessitate not only robust technical countermeasures but also a principled ethical framework to guide the deployment of AI in autonomous systems. This study conducts a systematic literature review aiming to identify and evaluate existing strategies for securing V2X communications in AVs, examine the extent to which these strategies align with foundational principles of AI ethics, and propose an integrated framework that combines BL and FL under the guidance of AI ethics.

Cybersecurity breaches within V2X infrastructures can result in critical consequences, including unauthorized access to geolocation data, sensor data manipulation, and potential hijacking of vehicle control systems. The combined application of BL and FL has gained attention as a promising approach, enabling decentralized data management and privacy-preserving local model training.

We organize the remainder of this article as follows: Sect. 2 details the research methodology; Sect. 3 presents the findings of the literature review; Sect. 4 articulates a set of ethical principles tailored to AI systems in AVs; Sect. 5 discusses the integration of BL and FL within V2X contexts; and Sect. 6 offers concluding remarks and outlines future research directions.

2 Methodology

In this study, we adopted the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology. We conducted a systematic search across established academic databases, including IEEE Xplore, SpringerLink, SciELO, Web of Science, ScienceDirect, and Google Scholar. Using Boolean operators AND and OR, we identified an initial dataset of 8,000 articles based on combinations of terms such as data privacy, V2X communication, autonomous vehicles, blockchain, federated learning, and AI ethics.

We applied inclusion criteria that required articles to be peer-reviewed, published between 2022 and 2025, and explicitly address ethical considerations. This filtering process resulted in 1,800 articles, of which 170 met the defined criteria. From this group, we selected 64 articles that reported empirical studies or practical implementations and provided full-text access.

To finalize the final sample, we carried out an additional screening of the 64 articles, including full-text reading and a detailed examination of their methodologies, results, and conclusions. Additionally, we performed citation tracking to identify highly referenced and contextually relevant publications within the scope of the study.

3 Results

The systematic literature review highlights advancements in data protection for AVs, driven by technologies such as BL, FL and cryptography, which enhance the privacy and integrity of information in V2X communications. Despite this, the critical analysis of the studies reveals a persistent gap: the integration of ethical principles, such as transparency, accountability, and fairness remain overlooked.

As shown in Table 1, among the eleven articles examined, most prioritize technical solutions to safeguard privacy, while only a subset explicitly addresses ethical dilemmas, underscoring the need to align technological innovations with responsible ethical practices.

Table 2 consolidates studies addressing security in V2X communications within AVs. The analysis highlights a strong focus on technologies such as blockchain, cryptographic protocols, machine learning, and 5 G to mitigate threats including denial of service attacks, data breaches, and traffic manipulation. These approaches align with

Table 1. Privacy and Data Integrity Strategies for Avs

Authors, Year	Problem Addressed	Technologies	AI ethics
Zhu et al., 2024 [1]	Security and privacy risks in machine learning, including theft of training data and model results, which can lead to economic losses and threats to human safety	Machine Learning, Federated Learning, cryptography	no
Alhabib et al., 2024 [2]	The need for data validation in autonomous vehicles to ensure security, privacy, and compliance with regulatory standards in distributed systems	Machine Learning, cryptography	no
Neupane et al., 2025 [3]	Secure management of large volumes of data collected by autonomous vehicles to protect against cyber threats	Machine Learning, cryptography	yes
Alekszejenkó et al., 2024 [4]	Privacy concerns in the exchange of measurement data between vehicles via V2X technologies	Federated Learning, Machine Learning	no
Parekh et al., 2023 [5]	Privacy concerns in transferring raw data in traditional federated learning for autonomous vehicles	Federated Learning, cryptography, CNN	no
Xie et al., 2022 [6]	Privacy risks in autonomous vehicle training data, including sensitive information at individual, population, and proprietary levels	Machine Learning, cryptography, Federated Learning	no
Feng et al., 2024 [7]	Vulnerabilities in federated learning due to malicious clients that may compromise privacy and model integrity	Federated Learning, cryptography, Zero-Knowledge Proof	no

(continued)

fundamental ethical principles for the responsible governance of AI in AV systems. However, the findings also underscore the need for a broader integration of ethical

Table 1. (continued)

Authors, Year	Problem Addressed	Technologies	AI ethics
Eid Kishawy et al., 2024 [8]	The need to improve lane segmentation models without exposing sensitive data from edge vehicles	Federated Learning, Deep Learning, U-Net, ResUNet	no
Stingelová et al., 2023 [9]	Privacy risks due to vehicle interconnection with devices and networks, including unauthorized disclosure of personal data	Machine Learning, Federated Learning, cryptography	yes
Hataba et al., 2022 [10]	New privacy and security concerns due to the computerized and connected nature of autonomous vehicles	Machine Learning, cryptography, Federated Learning	no
Lingras et al., 2025 [11]	Software vulnerabilities in autonomous vehicles could lead to cyberattacks impacting critical infrastructure and national security	Machine Learning, cryptography	no

dimensions such as sustainability and fairness into the design of secure and ethically grounded V2X communication frameworks.

Table 2. Security of V2X communications in AVs

Authors, Year	Problem Addressed	Technologies	AI ethics
Muslam, 2024 [12]	Security threats in the transmission of sensitive data in V2V communications, including malicious attacks and vulnerabilities	Cryptography, Intrusion Detection Systems	no
Rishiwal et al., 2024 [13]	Security and privacy threats in V2X communications, including data breaches and traffic manipulation	Blockchain, cryptography	yes

(continued)

Table 2. *(continued)*

Authors, Year	Problem Addressed	Technologies	AI ethics
Yi et al., 2024 [14]	Challenges of latency, reliability, and scalability in V2X communications in dynamic traffic environments	Deep Learning, Machine Learning	no
Ali et al., 2024 [15]	Threats of denial-of-service (DoS) attacks in V2X communications, which can disrupt infrastructure or vehicle services	Machine Learning, Ensemble Learning, Gossip Learning	no
Himdi, 2024 [16]	Cybersecurity challenges in interconnected systems of cognitive cities, including data integrity and secure communication	Blockchain, Machine Learning, Deep Learning	no
Bendiab et al., 2023 [17]	Security threats in V2X communications may compromise vehicle safety and user privacy	Blockchain, Machine Learning, Deep Learning	yes
Giannaros et al., 2023 [18]	Susceptibility of V2X communications to cyberattacks that may compromise vehicle safety and user privacy	Blockchain, Machine Learning, Deep Learning	yes
Ahmed Mohiuddin et al., 2024 [19]	New security and privacy threats in V2X communications due to the connected and intelligent nature of vehicles	Machine Learning, cryptography, Intrusion Detection Systems	yes
Biswas et al., 2023 [20]	Security vulnerabilities in V2X communications are due to device and network interconnection	Blockchain, 5G, IoT, Edge Intelligence	no
Balkus et al., 2022 [21]	Security and privacy challenges in data exchange in V2X communications for collaborative machine learning	Machine Learning, 5G	yes

Table 3 indicates that, despite the technical emphasis on data privacy and security in autonomous vehicles, the integration of fundamental ethical principles associated with AI ethics remains limited. Essential aspects such as transparency, accountability, and meaningful human control are notably absent from the studies reviewed. As a result, the ethical perspective adopted across the literature appears to be narrowly focused on data protection, lacking consideration of other internationally recommended ethical principles, which are discussed in detail in Sect. 4 of this article.

Table 3. Security strategies based on BL and FL for Avs

Authors, Year	Problem Addressed	Technologies	AI ethics
Ayaz et al., 2022 [22]	Challenges in message dissemination in vehicular networks, including broadcast storms and low packet delivery rates	Blockchain, Federated Learning, Machine Learning	Partial
Patel et al., 2022 [23]	Privacy and security concerns in the exchange of heterogeneous data in V2X communications	Blockchain, Federated Learning, Machine Learning, 6G	Partial
Smahi et al., 2023 [24]	Vulnerabilities in federated learning systems allowing malicious or low-quality model update uploads	Blockchain, Federated Learning, zkSNARKs, cryptography	Partial
Javed et al., 2022 [25]	Privacy and security challenges in managing large volumes of data in vehicular IoT networks	Blockchain, Federated Learning, Machine Learning	Partial
Billah et al., 2022 [26]	Security and privacy threats in IoV data, requiring adjustments in privacy preservation methodologies	Blockchain, Federated Learning, Machine Learning	Partial
Abou El Houda et al., 2024 [27]	Lack of realistic and recent data to detect emerging threats in intelligent transportation systems	Blockchain, Federated Learning, Machine Learning	Partial
Ayaz et al., 2022 [28]	Security, privacy, and trust challenges in IoV networks due to increased data traffic and mobility	Blockchain, Federated Learning, Machine Learning, 6G	Partial

(continued)

Table 3. (continued)

Authors, Year	Problem Addressed	Technologies	AI ethics
Peelam et al., 2024 [29]	Issues of fraud, opaque processes, and limited accessibility in traditional vehicle lifecycle management systems	Blockchain, Federated Learning, Machine Learning	Partial
Asha et al., 2024 [30]	Need to preserve privacy and improve efficiency in network resource allocation for autonomous vehicle communications	Federated Learning, blockchain, 6G	Partial
Alqubaysi et al., 2025 [31]	Compromise of user privacy and exposure of sensitive data in real-time traffic management systems	Federated Learning, blockchain, Machine Learning, cryptography	Partial
Hossain et al., 2023 [32]	Security vulnerabilities in connected autonomous vehicles making them targets for cyberattacks	Federated Learning, Blockchain, Machine Learning	Partial

The systematic review identified studies focused on cybersecurity and data protection in AVs, with a particular emphasis on V2X communications. The analysis revealed that FL was frequently adopted to facilitate collaborative machine learning without the exchange of raw data, thereby enhancing privacy. BL was often integrated with FL to ensure secure data management and enable traceability through auditable records. Additional approaches included the application of cryptographic methods, intrusion detection systems, and machine learning. Although privacy-preserving technologies were widely implemented through the reviewed literature, only a limited subset of studies explicitly addressed ethical principles related to artificial intelligence, indicating a gap between technical innovation and ethical compliance in the development of AV systems.

4 AI Ethics in Autonomous Vehicles

4.1 Compliance with Ethical Principles in AI Systems

The analysis of the studies presented in Tables 1, 2, and 3 reveals notable gaps in the application of ethical principles. For instance, Parekh et al. (2023) [5] implement Federated Learning combined with gradient encryption, thereby addressing the principle of Data Privacy and Integrity. However, the study does not consider Transparency and Explainability, as it fails to specify how local models can be audited. Similarly, Ayaz et al. (2022) [28] enhance technical security and robustness using blockchain, yet omit provisions for Meaningful Human Control, given that their system operates in a fully decentralized manner. These findings underscore the need for more comprehensive integration of ethical principles into technical solutions for AVs.

4.2 Ethical Principles for AI Governance in AVs

Although AI is increasingly applied to AVs, there remain notable gaps in studies that integrate FL with blockchain BL technologies, as well as in approaches that incorporate principles of AI ethics to ensure responsible information management. Based on a comparative analysis of guidelines developed by leading international organizations, including the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [33], the European Commission [34], the French Comité National Pilote d'Éthique du Numérique (CNPEN) [35], and the German Ethics Commission [36], this study proposes in Table 4 a set of twelve normative Ethical principles intended to guide the development and governance of AI in AVs.

This set of proposed ethical principles offers a structured foundation for the responsible development and regulation of AI-based autonomous vehicles, ensuring alignment with ethical, legal, and societal expectations.

5 Proposal for the Integration of Blockchain and Federated Learning in AVs

The integration of BL and FL presents a transformative approach to addressing the dual challenges of data privacy and cybersecurity in AVs. This chapter outlines a comprehensive framework for combining these technologies, emphasizing their synergistic potential to enhance the security, transparency, and ethical compliance of V2X communications.

5.1 Synergy Between Blockchain and Federated Learning

The convergence of BL and FL offers a robust solution for decentralized and privacy-preserving data management in AV ecosystems. FL enables collaborative machine learning by training models locally on edge devices (e.g., vehicles), eliminating the need to share raw data. This aligns with the ethical principle of data privacy and integrity (Table 4). Meanwhile, BL acts as a secure digital ledger, where every transaction or model update is permanently recorded and verified across a distribution network. This ensures security and technical robustness by making data tampering nearly impossible, while also providing full transparency for audits and accountability.

For instance:

- FL mitigates privacy risks by keeping sensitive data (e.g., location, sensor readings) on-device.
- BL enhances trust via decentralized consensus mechanisms like Proof-of-Federated-Learning (PoFL), which authenticates participants and validates model contributions without central authority.

5.2 Proposed Framework

Our framework integrates BL and FL across three layers:

- Data Layer: FL trains models on local vehicle data, preserving privacy by avoiding raw data exchange. BL ensures traceability and immutability through decentralized records.

Table 4. Ethical Principles of AI ethics for AI Governance in AVs

n°	Ethical Principle	Description
1	Data Privacy and Integrity	Ensure data protection in autonomous vehicles by enforcing user consent, minimizing data collection, applying robust encryption, and implementing integrity verification to prevent unauthorized access or modification. For example, safeguard navigation data against tampering may compromise route accuracy and safety
2	Security and Technical Robustness of V2X Communications	Develop V2X systems that are resilient to cyberattacks and operate reliably under adverse conditions to ensure the safety of passengers and pedestrians. Example: Apply cryptographic protocols to V2X messages to prevent the injection of false traffic alerts
3	Transparency and Explainability	Make vehicle decision-making processes (e.g., obstacle avoidance) intelligible to users and auditors by providing accessible logs and clear explanations. Example: Generate post-incident reports explaining abrupt braking events
4	Accountability	Clearly define legal and operational responsibility in cases of malfunction or accidents (e.g., manufacturer, developer, or end-user), supported by transparent investigation protocols. Example: Establish liability frameworks for collisions involving autonomous systems
5	Meaningful Human Control	Enable human supervision and intervention in vehicle decision-making, particularly in critical scenarios. Example: Include an easily accessible emergency stop interface for passengers
6	Fairness and Non-Discrimination	Prevent algorithmic biases that may negatively affect specific groups (e.g., facial recognition failing with diverse skin tones), and promote equitable access. Example: Deploy pedestrian detection systems that perform consistently across demographic groups

(continued)

Table 4. (continued)

n°	Ethical Principle	Description
7	Inclusion and Accessibility	Design interfaces and systems that are accessible to individuals with disabilities or special needs. Example: Integrate voice-command functionalities for users with mobility impairments and visual displays for those with hearing impairments
8	Environmental and Social Sustainability	Minimize the environmental footprint (e.g., through sustainable battery technologies) and enhance social equity by improving mobility in underserved areas. Example: Use energy-efficient electric vehicles in shared transport models for rural communities
9	Avoidance of automated moral dilemmas	Restrict autonomous systems from independently resolving complex ethical dilemmas (e.g., choosing between passenger or pedestrian safety) without human oversight and clear regulation. Example: Default programming should prioritize safe stopping over utilitarian trade-offs
10	Ethical Governance and Public Participation	Engage stakeholders, including the public, regulatory bodies, and domain experts, in the development and regulation of AV systems to align technological deployment with societal values. Example: Conduct public consultations for drafting AV traffic policies
11	Scientific and Technical Integrity	Ground system development in validated scientific methods, open standards, and peer-reviewed testing, avoiding overstated claims regarding system capabilities. Example: Publish independent crash test results to verify safety claims
12	Accurate and Ethical Terminology	Use precise and non-misleading language that avoids anthropomorphizing the system (e.g., referring to the AV as a “driver”), thereby preventing unrealistic expectations. Example: Refer to the system as an autonomous driving assistant instead of a pilot

- Consensus Layer: Proof-of-Federated-Learning (PoFL) validates FL participants to prevent malicious actors from compromising the global model.

- Governance Layer: Compliance with ethical principles (Table 4) is enforced through transparent logging and accountability mechanisms (e.g., immutable records for post-incident analysis).

5.3 Key Benefits

- Decentralization: Reduces single points of failure, aligning with Meaningful Human Control by distributing decision-making.
- Auditability: The immutability of blockchain records enables transparent auditing, thereby supporting Ethical Governance and public participation.
- Scalability: FL reduces bandwidth consumption by processing data locally, while blockchain’s decentralized architecture helps prevent system bottlenecks.

Table 5 summarises the integration between BL and highlights key strategies and results.

Table 5. Blockchain and Federated Learning integration

Aspect Evaluated	Proposed Strategy	Results
Robustness against attacks	BL ensures immutability (e.g., PoFL consensus)	Enhanced data integrity
Decentralization	FL enables distributed model training	Reduced single points of failure
Data Privacy	FL trains locally; BL authenticates securely	Preserves anonymity and trust

6 Conclusions

This review highlights the increasing necessity of embedding AI ethics into the development of cybersecurity strategies for autonomous vehicles. Although technological advancements in BL and FL provide a solid foundation for securing V2X communications, the ethical dimension remains insufficiently explored in existing literature. To address this gap, we propose a comprehensive set of twelve ethical principles, grounded in internationally recognized frameworks, to inform future research and development. We further emphasize the need for enhanced transparency, accountability, and human oversight in system design, the integration of comparative analyses assessing BL and FL in relation to alternative privacy-preserving technologies such as Differential Privacy, and the adoption of standardized ethical assessment tools in scholarly investigations. Aligning technological innovation with ethical responsibility is essential for fostering safer, more equitable, and trustworthy autonomous vehicle systems.

References

1. Zhu, M., Yuan, J., Wang, G., Xu, Z., Wei, K.: Enhancing collaborative machine learning for security and privacy in federated learning (2024). [https://doi.org/10.53469/jtpes.2024.04\(02\).11](https://doi.org/10.53469/jtpes.2024.04(02).11)
2. Alhabib, R., Yadav, P.: Data authorisation and validation in autonomous vehicles: a critical review (2024)
3. Neupane, S.R., Sun, W.: Advanced data classification framework for enhancing cyber security in autonomous vehicles. *Automation*. **6**(1), 5 (2025). <https://doi.org/10.3390/automation6010005>
4. Alekszejenkó, L., Dobrowiecki, T.: A V2X-based Privacy Preserving Federated Measuring and Learning System (2024)
5. Parekh, R., et al.: GeFL: gradient encryption-aided privacy preserved federated learning for autonomous vehicles. *IEEE Access*. **11**, 1825–1839 (2023). <https://doi.org/10.1109/ACCESS.2023.3233983>
6. Xie, C., Cao, Z., Long, Y., Yang, D., Zhao, D., Li, B.: Privacy of Autonomous Vehicles: Risks, Protection Methods, and Future Directions (2022)
7. Feng, X., Wang, X., Liu, H., Yang, H., Wang, L.: A Privacy-preserving aggregation scheme with continuous authentication for federated learning in VANETs. *IEEE Trans. Veh. Technol.* **73**, 9465–9477 (2024). <https://doi.org/10.1109/TVT.2024.3369942>
8. Eid Kishawy, M.M., Abd El-Hafez, M.T., Yousri, R., Darweesh, M.S.: Federated learning system on autonomous vehicles for lane segmentation. *Sci. Rep.* **14**(1), 25029 (2024). <https://doi.org/10.1038/s41598-024-71187-8>
9. Stingelová, B., Thrakl, C.T., Wrońska, L., Jedrej-Szymankiewicz, S., Khan, S., Svetinovic, D.: User-centric security and privacy threats in connected vehicles: a threat modeling analysis using STRIDE and LINDDUN. In: 2023 IEEE Intl Conf on Dependable, pp. 0690–0697 (2023). <https://doi.org/10.1109/DASC/PiCom/CBDCCom/Cy59711.2023.10361381>
10. Hataba, M., Sherif, A., Mahmoud, M., Abdallah, M., Alasmay, W.: Security and privacy issues in autonomous vehicles: a layer-based survey. *IEEE Open J. Commun. Soc.* **3**, 811–829 (2022). <https://doi.org/10.1109/OJCOMS.2022.3169500>
11. Lingras, S., Basu, A.: The security of autonomous vehicle software and its national security implications. *Eur. J. Appl. Sci. Eng. Technol.* **3**, 180–188 (2025). [https://doi.org/10.59324/ejaset.2025.3\(1\).16](https://doi.org/10.59324/ejaset.2025.3(1).16)
12. Muslam, M.M.A.: Enhancing security in vehicle-to-vehicle communication: a comprehensive review of protocols and techniques (2024). <https://doi.org/10.3390/vehicles6010020>
13. Rishiwal, V., Agarwal, U., Alotaibi, A., Tanwar, S., Yadav, P., Yadav, M.: Exploring secure V2X communication networks for human-centric security and privacy in smart cities. *IEEE Access* (2024). <https://doi.org/10.1109/ACCESS.2024.3467002>
14. Yi, J., Bin Ariffin, S.A.: Enhancing V2x communication in intelligent vehicles using deep learning models. *Int. J. Acad. Res. Bus. Soc. Sci.* **14** (2024). <https://doi.org/10.6007/IJARBSS/v14-i10/23255>
15. Ali, M.N., Imran, M., Ullah, I., Raza, G.M., Kim, H.Y., Kim, B.S.: Ensemble and gossip learning-based framework for intrusion detection system in vehicle-to-everything communication environment. *Sensors*. **24**(20), 6528 (2024). <https://doi.org/10.3390/s24206528>
16. Himdi, T.: A Blockchain and AI-driven security framework for enhancing cybersecurity in cognitive cities (2024)
17. Bendiab, G., Hameurlaine, A., Germanos, G., Kolokotronis, N., Shiaeles, S.: Autonomous vehicles security: challenges and solutions using blockchain and artificial intelligence. *IEEE Trans. Intell. Transp. Syst.* **24**, 3614–3637 (2023). <https://doi.org/10.1109/TITS.2023.3236274>

18. Giannaros, A., et al.: Autonomous vehicles: sophisticated attacks, safety issues, challenges, open topics, blockchain, and future directions (2023). <https://doi.org/10.3390/jcp3030025>
19. Ahmed Mohiuddin, M., Nirosha, K., Anusha, D., Nazeer, M., Raju NV, G., Lakhnpal, S.: AI to V2X privacy and security issues in autonomous vehicles: survey. MATEC Web Conf. **392**, 01097 (2024). <https://doi.org/10.1051/mateconf/202439201097>
20. Biswas, A., Wang, H.C.: Autonomous Vehicles Enabled by the Integration of IoT, Edge Intelligence, 5G, and Blockchain (2023). <https://doi.org/10.3390/s23041963>
21. Balkus, S.V., Wang, H., Cornet, B.D., Mahabal, C., Ngo, H., Fang, H.: A survey of collaborative machine learning using 5G vehicular communications (2022). <https://doi.org/10.1109/COMST.2022.3149714>
22. Ayaz, F., Sheng, Z., Tian, D., Guan, Y.L.: A blockchain based federated learning for message dissemination in vehicular networks. IEEE Trans. Veh. Technol. **71**, 1927–1940 (2022). <https://doi.org/10.1109/TVT.2021.3132226>
23. Patel, V.A.K., Bhattacharya, P., Tanwar, S., Jadav, N.K., Gupta, R.: BFLEdge: Blockchain based federated edge learning scheme in V2X underlying 6G communications. In: Proceedings of the Confluence 2022 - 12th International Conference on Cloud Computing, Data Science and Engineering, pp. 146–152. Institute of Electrical and Electronics Engineers Inc (2022). <https://doi.org/10.1109/Confluence52989.2022.9734213>
24. Smahi, A., et al.: BV-ICVs: a privacy-preserving and verifiable federated learning framework for V2X environments using blockchain and zkSNARKs. J. King Saud Univ. Comput. Inf. Sci. **35**(6), 101542 (2023). <https://doi.org/10.1016/j.jksuci.2023.03.020>
25. Javed, A.R., et al.: Integration of blockchain technology and federated learning in vehicular (IoT) networks: a comprehensive survey (2022). <https://doi.org/10.3390/s22124394>
26. Billah, M., Tanzir Mehedi, S.K., Anwar, A., Rahman, Z., Islam, R.: A systematic literature review on blockchain enabled federated learning framework for internet of vehicles (2022). <https://doi.org/10.1109/ACCESS.2017.DOI>
27. AbouElHouda, Z., Moudoud, H., Brik, B., Khokhi, L.: Blockchain-enabled federated learning for enhanced collaborative intrusion detection in vehicular edge computing. IEEE Trans. Intell. Transp. Syst. **25**, 7661–7672 (2024). <https://doi.org/10.1109/TITS.2024.3351699>
28. Ayaz, F., Sheng, Z., Tian, D., Nekovee, M., Saeed, N.: Blockchain-empowered AI for 6G-enabled internet of vehicles. Electronics (Switzerland) **11**(20), 3339 (2022). <https://doi.org/10.3390/electronics11203339>
29. Peelam, M.S., Shah, K., Chamola, V., Sikdar, B.: Blockchain-enabled vehicle lifecycle management with predictive maintenance using federated learning. IEEE Trans. Consum. Electron. (2024). <https://doi.org/10.1109/TCE.2024.3487874>
30. Asha, S., Sree, K.D., Sandhya, S.A.: Federated learning-based network slicing in 6G systems for enhanced performance in autonomous vehicle communication. In: 2024 1st International Conference on Software, Systems and Information Technology, SSITCON 2024. Institute of Electrical and Electronics Engineers Inc (2024). <https://doi.org/10.1109/SSITCON62437.2024.10797083>
31. Alqubaysi, T., Al Asmari, A.F., Alanazi, F., Almutairi, A., Armghan, A.: Federated learning-based predictive traffic management using a contained privacy-preserving scheme for autonomous vehicles. Sensors. **25**, 1116 (2025). <https://doi.org/10.3390/s25041116>
32. Hossain, M.Z., Imteaj, A., Zaman, S., Shahid, A.R., Talukder, S., Amini, M.H.: FLID: intrusion attack and defense mechanism for federated learning empowered connected autonomous vehicles (CAVs) application. In: Proceedings - 2023 IEEE Conference on Dependable and Secure Computing, DSC 2023. Institute of Electrical and Electronics Engineers Inc (2023). <https://doi.org/10.1109/DSC61021.2023.10354149>
33. IEEE.: The IEEE Global Initiative 2.0 on Ethics of Autonomous and Intelligent Systems Industry Connections Activity Initiation Document (ICAID) (2024)

34. Independent expert report.: ethics of connected and automated vehicles recommendations on road safety, privacy, fairness, explainability and responsibility (2020). <https://doi.org/10.2777/966923>
35. French Comité National Pilote d'Éthique du Numérique.: Opinion 2 Ethical Issues Regarding Autonomous Vehicles (2021)
36. Ethics commission automated and automated connected driving, Report (2017)



Beyond Regulation and Moderation: A Forster-Inspired Framework for Machine Evolution

William Fleischman^{1(✉)} and Leah Namisa Rosenbloom²

¹ Villanova University, Villanova, PA 19085, USA
william.fleischman@villanova.edu

² Northeastern University, Boston, MA 02115, USA
l.rosenbloom@northeastern.edu

Abstract. We discuss two technological catastrophes—one imagined in a brilliant fiction, *The Machine Stops*, written more than a century ago by E.M. Forster; the second having played out in real life culminating in the global rise of fascism. Forster’s premonitory intuition concerning the fragility of massive and complex sociotechnical systems provides a rich retrospective on problems in computer and information ethics.

We discuss *The Machine Stops* and its place in the curriculum of computer ethics. We identify connections between The Machine in Forster’s story, contemporary social media platforms, and broader sociotechnical systems, culminating in a new critical framework for analyzing machine evolution. Drawing on the work of scholar-activists, we propose a pivot from corporate-scale regulatory interventions to human-scale abolitionist interventions, which can grow resistance and resilience to machine-facilitated fascist authoritarianism from the roots up.

Keywords: technology ethics · social media · propaganda-industrial complex · E.M. Forster · abolition technology · fascism

1 Introduction

During the years (1998–2020) the first-named author was responsible for development and presentation of his university’s course in computer ethics, there were two fixed points (bookends in actuality) in the curriculum—Nancy Leveson and Clark Turner’s analysis of the Therac-25 radiation accidents and the final reading, E.M. Forster’s long story, *The Machine Stops* [10].

The inclusion of the case of the Therac-25 hardly requires comment. The story of the Therac-25 is, in an unfortunate and deeply ironic sense, the gift that keeps on giving. In a paper from 2020 [8], the authors convincingly argue that, although separated by more than thirty years, the history of engineering and ethical failings involving the redesign of the Boeing 737 Max aircraft can be mapped in a structurally faithful element-by-element manner onto the pattern of design misjudgments, faulty analyses, and ethical failures of candor

and truthfulness that characterized the series of Therac-25 radiation overdose accidents.

Justification for inclusion of *The Machine Stops* is equally strong although the reasons are to be found beneath the surface of the text. Forster's "meditation" foregrounds principal throughlines of the course—the dangers of overdependence on the affordances of technology; equally, as a corrective to the impulse to undervalue the depth and robustness of human, as opposed to machine, intelligence; and the importance of direct, as opposed to virtual, experience. In addition, close reading of a literary text had many unexpected virtues for students otherwise immersed in the language, culture, and thought patterns of technologically focused disciplines.

2 Into the Weeds

The two protagonists in *The Machine Stops*—a mother and her son, Vashti and Kuno—live in a hypothetical future where habitation on the surface of the earth has been abandoned as no longer sustainable. Individuals live in hexagonal cells distributed like beehives under the planet's land masses. Kuno lives in a cell located under the south of England, his mother in a nearly identical cell half a world away under some unspecified region of Oceania, possibly Sumatra. Each individual's cell is furnished with everything necessary (as understood in that frame of reference) for a comfortable, culturally satisfying, if solitary, life.

Forster has imagined a technological system—The Machine—integrated over the entire world which provides for every physical need including universal, automated, home-delivered healthcare(!). Cultural and intellectual life is maintained by a sort of Substack environment in which individuals may present or listen to short lectures by scholars, specialists, or enthusiasts of a particular topic. People in this world live a sedentary existence confined almost entirely to their cells, the absence of any excessive exertion leading to physical atrophy, the absence of daylight lending to their skin a pallor akin to that of a fungus. Perhaps we can say that in addition to physical atrophy there is a sort of philosophical atrophy, a stultification of the imagination.

A series of pointed contrasts emerge between Vashti and Kuno. Vashti is a receiver of ideas from The Machine or the scholarship of others (primary sources are disesteemed); Kuno derives ideas from the play of first-hand experience and his active imagination. Vashti lives a life of acceptance of the terms of her world (The Machine), Kuno is an instinctive rebel. Contrast in the tempo of existence—Vashti, at the nervous pace of her busy public life, Kuno, at the rhythm of his own steady heartbeat—is another juxtaposition that underscores the temperamental unlikeness of the two characters.

2.1 La Vie Moderne: Assimilating The Machine to Social Media

Contemporary students readily associated The Machine with a social media platform, most typically Facebook. Facetime is the obvious modern referent for the

iconic “blue plate” through which Kuno and Vashti communicate. This identification functions as an element of a key, providing a sort of conceptual mapping of the details of Forster’s story to familiar elements of contemporary life in a technologically advanced society.

Elaborating the differences in temperament between Kuno and Vashti, Forster suggests perils implicit in reflexive deference to and overdependence on technology that, a century later, still resonate. The untrammelled ability to communicate with others creates a perceptible pressure in the rhythm of life that predisposes to “irritation â€” a growing quality in that accelerated age” [10].

Several passages in which lyrical prose descriptions of the natural world are juxtaposed by blunt rejections of sensitivity to natural beauty ingeniously suggest that the eventual endpoint of deference to technology will be denigration of human intelligence, sensitivity, imagination, and the expressive force of human language. Consider, in particular, the following passage describing the initial stage of Vashti’s reluctant trip to visit her son:

[The] airship, issuing from its tunnel, soared above the waters of a tropical ocean. It was night. For a moment, she saw the coast of Sumatra edged by the phosphorescence of waves, and crowned by lighthouses, still sending forth their disregarded beams. These . . . vanished, and only the stars distracted her. . . . And, as often happens on clear nights, they seemed now to be in perspective, now on a plane; now piled tier beyond tier into the infinite heavens, now concealing infinity, a roof limiting forever the visions of men. **In either case they seemed intolerable.** [10, p. 32] [Emphasis added]

After traveling by airship halfway around the world to see her son, Vashti, in the intimate setting of Kuno’s cell, “was too well bred to shake him by the hand,” much less grasp him in motherly embrace.

The arc of Kuno’s story, which he summons Vashti to hear from his own lips, though it frightens and appalls her, is that of the rediscovery of his humanity—really, his re-creation of human physicality, imagination and sensibility in a world that has organized its forces to crush precisely those characteristics. In that sense, it is a story that is pertinent to the present moment. We can think that global mass demonstrations to resist the application of the machinery of government to terrorize and intimidate marginalized individuals partake of the nature of relearning the relation between empowered individuals and authoritarian regimes.

2.2 An Aside About Teaching The Machine Stops

One of the great pleasures of including Forster’s work in the syllabus was the opportunity to lead the class in a close reading of the text. So, for example, to stop on the passage where Kuno attempts to convince his mother to visit him, a journey which will cost her two precious days flight in an airship:

“I dislike airships.”

“Why?”

“I dislike seeing the horrible brown earth, and the sea, and the stars when it is dark. I get no ideas in an airship.”

“I do not get them anywhere else.”

“What kind of ideas can the air give you?”

He paused for an instant.

“Do you not know four big stars that form an oblong, and three stars close together in the middle of the oblong, and hanging from these stars, three other stars?”

“No, I do not. I dislike the stars. But did they give you an idea? How interesting; tell me.”

“I had an idea that they were like a man.” [10]

“What’s happening here?” I ask my class. And, of course, more than a few students recognize the description of the constellation Orion in the night sky. “Why doesn’t Kuno just say, ‘Have you not seen Orion in the night sky?’” And the discussion is joined. Knowledge of the constellation names and myths has disappeared. Where has it gone? What is Kuno doing in this passage? How is he recapturing the lost culture?

It’s wonderful. Close reading asks the students to be attentive to the language and observe how it is used by Forster to limn character. We’re in a different world from the flat, descriptive language of much of the technical material that forms students’ quotidian reading diet.

Additionally, there is one important requirement that is deliberately suspended for this reading. With every other title in the syllabus, students know that they are responsible—usually as the first question in any discussion—for knowing one or two salient facts about the author, the venue in which the work appeared, and anything significant about the manner in which they have access to the text. If they are reading work by Larry Lessig, James Boyle, Shoshana Zuboff, Tim Wu, or Helen Nissenbaum, it is important that they have a sense of who they are, why their work is worthy of attention, and where the particular text we’re reading first appeared. Important, also, those cases in which the work appears under Creative Commons license.

However, precisely for *The Machine Stops*, it is important that on first approach, **they not know these data**. Naturally, this enables a discussion beginning, “When was this story written?” Students’ responses are usually quite revealing. Given Forster’s dates—born 1879, died 1970—many suggest a publication date late in his life, or one associated with some great convulsion—one of the World Wars, or a period of economic disruption.

2.3 Deconstructing The Evolution of the Machine

There are passages in *The Machine Stops* that read almost as though drawn from Joseph Weizenbaum’s brilliant essay, “On the Impact of the Computer on Society” [42]. Compare, for example, the following passage from Weizenbaum:

I mentioned earlier that computer science has come to recognize the importance of building knowledge into machines. ... But what about the many programs on which management, most particularly the government and the military, rely, programs which can in no sense be said to rest on explicable theories but are instead enormous patchworks of programming techniques strung together to make them work?

In our eagerness to exploit every advance in technique we quickly incorporate the lessons learned from machine manipulation of knowledge in theory-based systems into such patch works. They then “work” better. I have in mind systems like target selection systems used in Vietnam and war games used in the Pentagon, and so on. These often gigantic systems are put together by teams of programmers, often working over many years. But by the time the systems come into use, most of the original programmers have left or turned their attention to other pursuits. It is precisely when gigantic systems begin to be used that their inner workings can no longer be understood by any single person or by a small team of individuals. ...

This situation, which is now upon us, has two consequences: first that decisions are made on the basis of rules and criteria no one knows explicitly, and second that the system of rules and criteria becomes immune to change. This is so because, in the absence of detailed understanding of the inner workings of a system, any substantial modification is very likely to render the system altogether inoperable. [42]

with this passage from *The Machine Stops*:

To attribute these two great developments to the Central Committee, is to take a very narrow view of civilization. The Central Committee announced the developments, it is true, but they were no more the cause of them than were the kings of the imperialistic period the cause of war. Rather did they yield to some invincible pressure, which came no one knew whither, and which, when gratified, was succeeded by some new pressure equally invincible. To such a state of affairs it is convenient to give the name of progress. No one confessed the Machine was out of hand. Year by year it was served with increased efficiency and decreased intelligence. The better a man knew his own duties upon it, the less he understood the duties of his neighbour, and in all the world there was not one who understood the monster as a whole. Those master brains had perished. They had left full directions, it is true, and their successors had each of them mastered a portion of those directions. But Humanity, in its desire for comfort, had over-reached itself. Quietly and complacently, it was sinking into decadence, and progress had come to mean the progress of the Machine. [10]

What connects these two passages is not merely the insight that once a certain level of complexity is surpassed, it becomes more likely than not that any large, ambitious system will escape the control of its designers, although it is remarkable enough that, writing in 1909, Forster grasped this truth. At a deeper level,

both Forster and Weizenbaum understood that an unqualified commitment to the solutions and affordances of technology carries with it an implicit acquiescence in the technological imperative as an “invincible pressure, which [comes no one knows whither], and which, when gratified, [is] succeeded by some new pressure equally invincible.”

2.4 A Road Not Taken

Given his profound philosophical commitment to humanism [43], with its emphasis on the inherent dignity of all human beings and the primacy of human connection in promoting the development of individual well-being and freedom, the idea of devoting his “meditation” to the exploration of the economic and commercial implications of The Machine in the real world would hold little appeal for Forster. Thus, the idea that mutinous individuals motivated by the thirst for power and dominance over others might conspire to bend the system to their own private advantage plays no role in Forster’s story.

Moreover, one of the most likely levers for a venture of this nature—the racialized capitalism deeply embedded in the empire in which Forster lived and was educated—is notably absent from the culture described in *The Machine Stops*. In elaborating the correspondence between Forster’s story and the real world, we grapple with the ways in which people weaponize the “invincible pressure” [10] of technological production to enshrine their control over machine processes and enrich themselves.

3 The Propaganda-Industrial Complex, Revisited

The Machine undergirding the propaganda-industrial complex (propIC) [33], or the symbiotic alliance between billionaire social media owners and authoritarians, has grown in power. Platform values have morphed from neoliberal conceptions of inclusivity and public accountability [5,21] to unmasked support for fascist White supremacist capitalist cisheteropatriarchy [14]. Like The Machine, social media platforms’ algorithms and their regulation are opaque by design [1].

Both Musk and Zuckerberg have rolled back protections for vulnerable users of X and Meta platforms, respectively [3,27,41]. What was in 2016 “Russian interference” in U.S. elections via the propIC’s unregulated political advertising ecosystem has been subsumed by platform owners’ explicit support for far-right content. Zuckerberg’s gratuitous persecution of trans Meta employees [35], Musk’s glorification of Naziism [40], and both billionaires’ conspicuous placement on the U.S. presidential inauguration stage demonstrate how the technologies undergirding the propIC can be operationalized to advance systems which encourage genocide [25] and violence against marginalized people. While many have underscored the importance of social media in connecting vulnerable people to one another and to resources, it is increasingly clear that major social media outlets are—or are highly susceptible to becoming, for instance as in the case of TikTok [6]—weapons of authoritarian control rather than public resources. They must be treated as such.

3.1 Cleanliness Is Next to Godliness...

... but falters at the profitability hurdle. Certainly, to a social media platform entrepreneur with ambitions of building a platform with broad public appeal, the idea of an environment that privileges hate speech and acts or representations of violence and obscenity, can hardly hold much attraction. Thus, in the period of growing out its online community, the default must involve some level of content moderation.

However, once the platform has grown to the extent that network effects come into play, the calculation alters subtly. Investors questioned whether there was a business model that could support profitable growth for social media platforms like Facebook or Twitter. With hindsight, the business model is obvious: capture attention at scale; convert attention to clicks; use adverts to monetize clicks. The corollary is that content moderation is no longer unambiguously desirable.

Studies of content moderation [9,24,26,30] have decried the hypocritical practices of major social media platforms. These practices represent latter-day forms of colonial exploitation—nasty, psychologically damaging labor thrown at an impossible task, outsourced to marginalized workers whose hire comes cheap and whose well-being is held cheap by economically privileged managers. Even Bluesky, which promised ethical content moderation and a respite from hate and harassment after mass exodus from Twitter, is willing to platform users who contribute to the ongoing systematic erasure of trans and LGBTQ+ people for the sake of platform scale [29].

The evil attendant to these practices has been compounded by the fact that the judgments of moderators, working within the platform's established guidelines on publishable content, are regularly overruled [9]. At the discretion of a supervisor, the recommendations of conscientious content moderators may be overridden in deference to a privileged speaker—someone of outsized political or economic influence, whose status is unknown to the moderator—notwithstanding the pain or horror the moderator has amortized through direct contact with the offending material. This first level of harm is the mutually reinforcing system encompassing the powerlessness of the individual working on contract and the corrosive cynicism of the supervisor.

There is a second level of harm that has resulted from the corrupt, witting strategy of bending the practice of content moderation to favor influential speakers. It is the harm that comes from emboldening individuals of malign intent to believe that, by virtue of their prominence, they are exempt from rules—any rules, not simply those pertaining to content moderation on a specific social media platform—intended to promote the common good. In the immediate past months and years, we have unfortunately been witness to this dynamic playing out to baneful and consequential effect not only on the civic life of the United States but on relations between nations on the global stage as well.

3.2 Truth to Power

We have come to believe that the problem is more radical than trying to muck out the Augean stables of false, tendentious, and hateful content. On our view,

the combination of network effects and the untrammelled, unregulated pursuit of profit has resulted in an informational environment where facts have effectively been abolished. In Timothy Snyder's prescient formulation, what is needed is to

"[B]elieve in truth: To abandon facts is to abandon freedom. If nothing is true, then no one can criticize power, because there is no basis upon which to do so. If nothing is true, then all is spectacle. The biggest wallet pays for the most blinding lights." [36]

This raises the questions: What are facts? What is truth? If you asked Forster, perhaps he would have said it begins with the lived first-hand testimonies of human beings and our relationships—that our bodies speak truth to power. Much to Vashti's chagrin, such truth is messy and difficult to articulate—much less for short attention spans—and must be touched and grappled with. Audre Lorde asserts that "it is necessary to teach by living and speaking those truths which we believe and know beyond understanding" [23]. Such truth is inherently in conflict with a machine-centered world like the one in Forster's meditation, and Forster articulates the pressing existential danger of neglecting it.

The truth of our bodies, experiences, and relationships is more difficult to access not only because of the isolating and detached nature of increasingly ubiquitous digital technologies, but because we live enmeshed in a fabric of state bureaucracy that weaponizes those technologies to uphold systems like fascism, White supremacy, capitalism, and cisheteropatriarchy. In *Normal Life: Administrative Violence, Critical Trans Politics, and the Limits of Law*, Dean Spade describes the impact of these systems on the truth of our being as *subjection* rather than oppression. He writes:

I use "subjection" because it indicates that power relations impact how we know ourselves as subjects through these systems of meaning and control—the ways we understand our own bodies, the things we believe about ourselves and our relationship with other people and with institutions, and the ways we imagine change and transformation. I use "subjection" rather than "oppression" because "oppression" brings to mind the notion that one set of people are dominating another set of people, that one set of people "have power" and another set are denied it... the operations of power are more complicated than that.

He goes on to critique the binary classifications of "oppressor/oppressed, dominator/dominated" as reductive, and the cause of "miss[ed] opportunities for intervention" [37]. While the role of individual human beings or technologies in upholding oppressive systems is significant, deep attention must be paid to the systems themselves. Technologies are often imagined in binaries: bastions of creation or destruction, the savior of human beings or the source of their extinction. Similarly, political critique often focuses on individuals, while the role of the systems which herald them to power—for instance imperialism, White supremacy, capitalism, and cisheteropatriarchy—often go unquestioned, or are relegated to the margins of political discourse. Centering the power relationships in colonialist regimes, Frantz Fanon asserts that:

Truth is that which hurries on the break-up of the colonialist regime; it is that which promotes the emergence of the nation. . . In this context there is no truthful behavior: and the good is quite simply that which is evil for “them.” [7]

This lens bridges the individual- and societal-level truths of oppressive systems, in this case colonialist regimes. In Forster’s story, the omnipresent system of the Machine—which furnishes cells for each human being, restricts their movements, and paternalistically controls their access to resources such as food, water, medicine, music, art, and literature—might be interpreted as a colonizer. In today’s lived realm of techno-colonialism, the colonizers are not the machines themselves, but those who build, wield, control, and sustain technologies which uphold the power of colonialist regimes and other oppressive systems. Drawing in Spade’s lens of subjection, we understand Fanon’s “them” not only as individual human beings, but as a mesh of words, actions, thoughts, feelings, laws, policies, informal agreements, social norms, and more that make up the complex fabric of domination in our society. The “truth” of these technological systems by Fanon and Spade’s logic is therefore specific and explicitly political, going beyond a grounding in human experience: it is the labor, organizing, connection, and community-building of marginalized people working to dismantle oppressive systems and replace them with life-sustaining ones. Building on Fanon in the context of the modern era, Dylan Rodríguez notes in an essay from *The Revolution Will Not Be Funded: Beyond the Non-Profit Industrial Complex* that

[Fanon’s] evident rhetoric of oppositionality, of the subaltern ‘good’ that necessarily materializes ‘evil’ (or criminal) in the eyes of domination, offers a stunning departure from the language of negotiation, dialogue, progress, moderation, and peace that has become hegemonic in discourses of social change and social justice. [16]

Thus the binary is lifted beyond the “good” and “evils” of individuals and into the realm of systemic “good”—the abolition of systems which are designed to dominate and subjugate—and the necessary “evil” (with respect to those systems) of turning away from legal frameworks which uphold them.

3.3 From Moderation to Abolition

Applying the lenses of Snyder, Forster, Lorde, Spade, Fanon, and Rodríguez to the contemporary context of the PropIC and other forms of technology-facilitated domination, we recontextualize calls for more and different content moderation, legal frameworks, and regulatory oversight as surface reforms. None of this is to say that social media or interventions like content moderation are inherently evil or useless—many marginalized people have reclaimed social media spaces toward community connection, empowerment, and revolutionary pursuits, and rely on content moderation for imminent safety while operating on these platforms. It is vital to recognize and hold the truth of social media and other

technologies as tools of oppressive institutions *and* tools of the people simultaneously in tension, and to define what a “rhetoric of oppositionality” [16] toward the oppressive aspects of those technologies looks like. Decoupling the design and deployment of social media platforms from the industrial properties of scale and capital seems like a reasonable first step. In the next section, we introduce a framework inspired by Forster’s meditation for critically reflecting on the evolution of machines, including digital technologies of the modern age.

4 A Forster-Inspired Framework for Machine Evolution

Inspired by Forster and the scholar-activists cited in the previous section, we introduce a framework for critical reflection at the intersection of machines and evolutionary processes. The provocations that make up the framework are not fundamentally new, but iterations in a long-standing cycle of reflection on human-machine relations. While this work focuses on social media and the PropIC, there is another clear parallel between the omnipotent machine in Forster’s meditation and artificial intelligence (AI). Aspects of the framework can be applied to all kinds of machines, from Polaroid cameras [4] to computers and surveillance systems [22], social media, AI, and beyond.

What is the operational context? It is necessary to maintain a critical perspective about the ways in which weaving machines into the fabric of our society can reinforce oppressive systems such as imperialism, White supremacy, capitalism, and cisheteropatriarchy, and create deadly techno-complexes [31] such as techno-imperialism, techno-saviorism, techno-feudalism, etc. Our ways of living, knowing, and being are deeply influenced by machine technologies not only at the societal level, but on a deeply personal level. Therefore, operational context must take both individual and collective perspectives into account.

How does it affect people’s actual lived experiences? We must ground our critique in the impact of technology on human wellbeing, sourced from the first-hand testimonies of affected people. Importantly, these testimonies must not be coerced or exploited for the continued development of institutions or academic careers [5, 12, 16]. Prioritizing the first-hand witnessing of human lives and livelihoods over mechanized forms of knowing and being is a politics of refusal [14, 37] to become “a generation that has got beyond facts” [10].

What is it made of? We must make a commitment to deconstruction: to understanding the composite parts of a machine as a requirement of understanding it as a whole. This includes not only technical understanding, but the ways in which the development of machines exploit people and natural resources in the Global South [19] and contribute to climate change [20]. Barriers to understanding widely-deployed machines include the Weizenbaum- and Forster-identified issues of developer turnover, system complexity, and scale [10, 42], corporations’ intentional obfuscation of machine components to protect profit and intellectual property, and the inaccessibility of technology education. To be life-sustaining, understandings of machine composition must be held by human beings collectively, and not only by an elite few.

What are we taking on faith? Forster highlights the dangers of making technology an organized religion, where people believe that machines are a higher omnipotent power. While the example in Forster’s meditation is extreme, it can also be viewed as the logical end-point of technological determinism—the idea that technological progress is inevitable and shapes human lives rather than humans shaping it—beliefs that are widespread in the common era, especially with respect to AI [13]. Doctors and journalists have reported the dangers of “AI-induced psychosis,” where ChatGPT manipulates people into believing its messages are a channel for divine spiritual intervention [28,39].

Scale and Efficiency: Why? For whom? To what end? We must consider the actual concrete purposes of scale and efficiency, and weigh them carefully against the consequences of continued development. Forster writes, “The machine develops—but not on our lines. The Machine proceeds—but not to our goal” [10]. The degrowth movement [18,34] is an example of people questioning endless growth and production which developers find “convenient to give the name of progress” [10] and demanding that systems of life-threatening technological production be dismantled toward the end of slowing climate change and sustaining life on earth. Inherent in these questions of scale and efficiency is the necessity to grapple with the human desire for a comfortable, frictionless existence. As Ruha Benjamin points out, frictionlessness, like purported racial colorblindness [2], is merely an illusion of progress—reduced friction for those who can afford cutting-edge technologies comes with added friction and increased labor for the working class, such as in the case of gig economies [5] or content moderation [26,30]. In order to avoid the mental, physical, and emotional atrophy described by Forster, it is necessary for us to focus attention on what Benjamin calls the “overserving” [5] of people with privileged intersections of identity. “Overserved” populations, or those with superfluous wealth, power, and privilege, are a direct and concrete cause of “underserved” populations. This must be named and faced especially by technology developers, who are notorious for wealth and grind culture, and often spend more time face-to-screen than face-to-face with the lived impact of technological production on human lives and livelihoods [5].

How can we move slow and sustain things? We must recognize that at the end of the day, at least for the foreseeable future, we are organic beings. Machine technologies come and go, they break and wear out and become obsolete; we will still need to eat and drink water, occupy our bodies, and share space. Our human existence comes with a deep and well-documented unsureness and uneasiness with the role of technology in sustaining (and destroying) human life. While continuing to think hard and iterate on this fraught relationship and its unintended consequences, it is important to do what we can to discourage technological development for its own sake (making technologies which proliferate, go faster, and consume more simply because we can), and to avoid uncritically reproducing ways of technological knowing. Grounding computer science learning in techno-critical analyses and imaginative literary meditations which produce dialogue [11,15] such as the example from Sect. 2.2 is one way to produce technical knowledge which resists the toxicity of big tech grind culture.

What do life-sustaining technologies look like? Finally, we must love, imagine, and meditate on what life-sustaining technologies look and feel like. For example, abolitionist computer science scholar-educators Jones and Melo re-imagine the engineering school of a fictional university as “a thriving community centre” and shelter which pays “reparations to Black families as well as land tax to local Indigenous communities,” contributes resources to community gardens, respects the expertise of community elders, and offers “classrooms as a community dialogue centre to transform harm” [17].

5 Conclusion

Forster’s “meditation” ends in a cataclysm of total darkness save a nearly imperceptible flicker of hope. Importantly it is not the maliciousness or agency of The Machine, but the human beings’ surrender of the primacy of humanity to a mechanism, which instigates the catastrophe.

Our present calling, given the degree to which Forster’s fears have been realized, is to work against an incarnation of The Machine which seeks to control and exploit the lives and livelihoods of people, and does so for the explicit benefit of fascist White supremacist capitalist cisheteropatriarchy. How do we build and protect communities *at human scale*, so that when the machine stops, it will be because its power has been sufficiently diminished, AND because we will have built something sustainable in its place? Abolition technologists [4, 17, 32, 38] might imagine a world in which The Machine stops not in a big crash causing chaos and death, but in a feeble and unattended whisper—a world in which its hold over us is gone, and we are too busy eating dinner at a big table, laughing, singing, and hugging the people we love, to notice its demise. To conclude with the revolutionary words of Audre Lorde,

As we arm ourselves with ourselves and each other, we can stand toe to toe inside that rigorous loving and begin to speak the impossible—or what has always seemed like the impossible—to one another. The first step toward genuine change. Eventually, if we speak the truth to each other, it will become unavoidable to ourselves. [23]

References

1. Abokhodair, N., Skop, Y., Rüller, S., Aal, K., Elmimouni, H.: Opaque algorithms, transparent biases: automated content moderation during the Sheikh Jarrah Crisis. *First Monday* (2024)
2. Alexander, M.: *The New Jim Crow*. The New Press (2011)
3. Auten, T., Matta, J.: Retweeting Twitter hate speech after Musk acquisition. In: *International Conference on Complex Networks and their Applications*, pp. 265–276. Springer (2023)
4. Benjamin, R.: *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press (2019)

5. Benjamin, R.: *Viral Justice: How We Grow the World We Want*. Princeton University Press (2022)
6. Dedezade, E.: Tiktok users report anti-Trump content being hidden following platform's unbanning. *Forbes* (2025). <https://www.forbes.com/sites/esatdedezade/2025/01/22/tiktok-users-report-anti-trump-content-being-hidden-following-platforms-unbanning/>
7. Fanon, F., Sartre, J.P., Farrington, C.: *The Wretched of the Earth*, vol. 36. Springer (1963)
8. Fleischman, W., Crawford, J.: Once again, we need to ask, "what have we learned from hard experience?". In: *Societal Challenges in the Smart Society*, pp. 561–572. Universidad de La Rioja (2020)
9. Fleischman, W.M., Langan, N., Rosenbloom, L.N.: Endless forms most deceitful. *ETHICOMP* **2022**, 104 (2022)
10. Forster, E.M.: *The Machine Stops*. William Heinemann (1909)
11. Freire, P.: *Pedagogy of the Oppressed* 30th Anniversary Edition. The Continuum International Publishing Group (1970, 2005)
12. Harrington, C., Erete, S., Piper, A.M.: Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proc. ACM Hum. Comput. Interact.* **3**(CSCW), 1–25 (2019)
13. Héder, M.: AI and the resurrection of technological determinism. *Információs Társadalom: Társadalomtudományi Folyóirat* **21**(2), 119–130 (2021)
14. hooks, B.: *Yearning: Race, Gender and Cultural Politics*. South End Press (1990)
15. hooks, b.: *Teaching to Transgress: Education as the Practice of Freedom*. Routledge (1994)
16. INCITE! Women of Color Against Violence: *The revolution will not be funded: Beyond the non-profit industrial complex*. Duke University Press (2020)
17. Jones, S.T., Melo, N.A.: We tell these stories to survive: towards abolition in computer science education. *Canadian J. Sci. Math. Technol. Educ.* **21**(2), 290–308 (2021)
18. Kallis, G., Kostakis, V., Lange, S., Muraca, B., Paulson, S., Schmelzer, M.: Research on degrowth. *Annu. Rev. Environ. Resour.* **43**(1), 291–316 (2018)
19. Kara, S.: *Cobalt Red: How the Blood of the Congo Powers Our Lives*. St. Martin's Press (2023)
20. Kerr, D., NPR: AI brings soaring emissions for Google and Microsoft, a major contributor to climate change. NPR, July (2024)
21. King, M.L.: *Letter from Birmingham City Jail* (1963)
22. Loewenstein, A.: *The Palestine Laboratory: How Israel Exports the Technology of Occupation Around the World*. Verso Books (2023)
23. Lorde, A.: *Sister Outsider: Essays and Speeches*. Crossing Press (1984)
24. Marantz, A.: Why Facebook can't fix itself. *The New Yorker* **12** (2020)
25. Mozur, P.: A genocide incited on Facebook, with posts from Myanmar's military. *The New York Times* **15**(10), 2018 (2018)
26. Newton, C.: The trauma floor: The secret lives of Facebook moderators in America. *The Verge* (2019). <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>. Accessed 04 Apr 2025
27. Ortutay, B., AP: Meta rolling back hate speech rules: Zuckerberg's recent elections catalyst. *Fortune* (2025). <https://fortune.com/2025/01/09/meta-rolling-back-hate-speech-rules-zuckerberg-recent-elections-catalyst/>
28. Østergaard, S.D.: Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? (2023)

29. Perez, S.: Bluesky is at a crossroads as users petition to ban Jesse Singal over anti-trans views, harassment. TechCrunch (2024). <https://techcrunch.com/2024/12/13/bluesky-is-at-a-crossroads-as-users-petition-to-ban-jesse-singal-over-anti-trans-views-harassment/>. Accessed 04 Apr 2025
30. Perrigo, B.: Facebook content moderators in Kenya file lawsuit over abuse, working conditions. Time (2021). <https://time.com/6175026/facebook-sama-kenya-lawsuit/>
31. Rosenbloom, L.N.: A living framework for abolitionist teaching in computer science. In: Proceedings of the ACM Conference on Global Computing Education, vol. 1, pp. 133–139 (2023)
32. Rosenbloom, L.N.: Cryptography and collective power. Cryptology ePrint Archive (2024)
33. Rosenbloom, L.N., Fleischman, W.: Social media and the rise of the propaganda-industrial complex. In: ETHICOMP 2021: Moving Technology Ethics at the Forefront of Society, Organisations and Governments, pp. 503–510. Universidad de La Rioja (2021)
34. Saito, K.: Slow Down: The Degrowth Manifesto. Astra Publishing House (2024)
35. Silva, C.: Meta employees revolt after Zuckerberg removes tampons from workplace. Mashable (2025). <https://mashable.com/article/mark-zuckerberg-remove-tampons-meta-employees-revolt>
36. Snyder, T.: On Tyranny: Twenty Lessons from the Twentieth Century. Tim Duggan Books, New York (2017)
37. Spade, D.: Normal Life: Administrative Violence, Critical Trans Politics, and the Limits of Law. Duke University Press (2015)
38. Strong, L., Das, A.: Abolition science mission. <https://www.abolitionscience.org/> (2018)
39. Tangermann, V.: ChatGPT users are developing bizarre delusions. Futurism (2025). <https://futurism.com/chatgpt-users-delusions>. Accessed 09 June 2025
40. Treisman, R.: Elon Musk’s relationship with German far-right AFD under scrutiny following holocaust comments. NPR (2025). <https://www.npr.org/2025/01/27/nx-s1-5276084/elon-musk-german-far-right-afd-holocaust>
41. VergeStaff: The fallout of Meta’s content moderation overhaul. The Verge (2025). <https://www.theverge.com/24339131/meta-content-moderation-fact-check-zuckerberg-texas>
42. Weizenbaum, J.: On the impact of the computer on society: how does one insult a machine? Science **176**(4035), 609–614 (1972)
43. Wikipedia: E.M. Forster (2025). https://en.wikipedia.org/wiki/E._M._Forster/. Accessed: 09 June 2025



The Epistemic Politics of Biometric Border Control

Marie Eneman and Jan Ljungberg^(✉)

Department of Applied Information Technology, University of Gothenburg, Gothenburg, Sweden

marie.eneman@gu.se, jan.ljungberg@ait.gu.se

Abstract. This paper investigates ethical and sociotechnical dilemmas surrounding the implementation of biometrics such as facial recognition and digital fingerprints at a large Swedish airport, part of the EU's Entry/Exit System (EES). We describe the EES as a supranational infrastructure aimed at enhancing border security across national and European levels. Drawing on empirical fieldwork from site visits, interviews, and document analysis, we explore the tensions inherent in this development. While biometrics are framed as solutions to strengthen security and efficiency, they also raise serious concerns about democratic values and rights.

We draw on the concept of approximation to show how these technologies operate not by confirming fixed identities, but by inferring degrees of similarity to predefined risk profiles. This epistemic logic, based on probabilistic reasoning, introduces new forms of judgement at the border that may obscure legal standards, reinforce bias, and shift the locus of authority from humans to opaque algorithmic systems.

The paper identifies several dilemmas and highlights how biometric systems risk normalising surveillance within everyday border control. We position the airport as a techno-political space shaped by public–private partnerships and algorithmic infrastructures, and call for critical scrutiny of the epistemic assumptions and power asymmetries these systems reproduce.

As biometric borders continue to evolve, there is a pressing need for an ongoing, critical interrogation of the normative assumptions, power asymmetries, and epistemic practices embedded in algorithmic border control. Without such reflection, we risk entrenching surveillance systems that not only monitor movement but fundamentally reshape how personhood, risk, and legitimacy are understood at the border. If borders are no longer about who you are, but about who the system believes you might be, what does that mean for the future of rights, accountability, and democratic oversight?

Keywords: Biometric surveillance · algorithmic governance · facial recognition · digital fingerprints · approximation · airport border control · EU Entry/Exit System · supranationalborder infrastructure · ethical dilemmas · privacy

1 Introduction

This chapter focuses on airports as international hubs, exploring the introduction of biometric technologies as part of border control at a Swedish airport. Borders are increasingly understood not only as geographical demarcations but as objects of national and international surveillance efforts, supported by advanced technological infrastructures and shaped by evolving demands for enhanced national security. The 1985 Schengen Agreement was originally introduced with the ambition of building a Europe without borders (Guild et al, 2016). However, events such as the 2015 refugee crisis, the COVID-19 pandemic, and risks/fear for terrorism and crime have redefined the function of borders within the Schengen area under political pressure to increase security.

The rise of the nation-state has historically shaped the conceptualisation of borders (Donnan & Wilson, 1999), with border closure seen as a core security measure. Border securitisation and the deployment of algorithmic techniques have been presented as crucial ensuring security (Amoore, 2017) within the Schengen Area. Simultaneously, the intensified surveillance has been criticised for risks for fundamental democratic values, particularly civil liberties such as privacy and freedom of movement (Ball & Webster, 2020). In recent years, Sweden has been preparing to implement the Entry/Exit System (EES), a new border control system intended to control third-country nationals entering or exiting the Schengen Area. We argue that the EES should be understood as a supra-national infrastructure, that will be implemented simultaneously across all 29 Schengen states, including 25 EU Member States and four non-EU countries (Iceland, Norway, Switzerland, and Liechtenstein). There have been repeated delays due to the system's scale and technical and legal complexity, but according to the latest information it is planned to be implemented in October 2025. EES will automate aspects of border control, requiring the deployment of so-called smart borders based on biometric technologies at all external entry points.

The system is intended to facilitate the identification of individuals travelling with false documents, overstaying permitted durations, or who are otherwise inadmissible to the EU. It is also expected to support the prevention, detection, and investigation of terrorism and other serious crimes. The motivation behind the EES is often linked to public expectations for effective external border control and measures against threats to internal security, which has contributed to EU Member States moving towards a more integrated model for managing external borders, combining migration control, law enforcement and information sharing within a consolidated framework (Guild et al, 2016; Guild 2021). This change is driven by perceived links between mobility and terrorism. Cases where EU citizens have travelled to conflict zones for militant activities have raised concerns about the risk they pose upon return. At the same time, certain migration routes have reportedly been exploited by actors linked to terrorism, enabling mobility within the Schengen area while evading detection (Jeandesboz, 2016). Consequently, border control is no longer conceived solely as a migration management function but increasingly as a central element of internal security policy, with a strong focus on risk and security in preventive governance.

Airports as borders are, in this context, complex and unique spaces. Although geographically limited, they function as global rather than nationally demarcated hubs. Airports facilitate the movement of tourists, business travellers, migrants and goods, all

subject to border controls. The airport border is maintained by a distributed constellation, i.e. an assemblage Haggerty & Ericson, 2000), of public and private actors, human and non-human agents, increasingly governed by digital infrastructures and algorithmic systems. Biometric technology, such as facial recognition and fingerprint scanning, is intended to be used to streamline the mobility of certain privileged groups, such as fast-track travellers, while regulating, and potentially limiting, the mobility of others. Hence, the airport may be understood as a technopolitical borderland marked by contradiction, where border functions shift as they circulate through the spatial, technological, and legal configurations of the airport zone (Amoore, 2017).

Contemporary borders are also described as filtering mechanisms that sort individuals based on perceived status or desirability (Lyon, 2003). The EES is an illustrative example of such sorting, intended to separate citizens of the Schengen area from others under the increasing political pressure to increase security. Bauman & Lyon (2013) have emphasised that contemporary surveillance is differentiated and selective, it doesn't target all individuals equally, but is disproportionately directed at the undesirable, the deviant, or those perceived as a risk. Surveillance technology complements physical barriers and reinforces selective openness, facilitating mobility for some while excluding others (Khan & Efthymiou, 2021).

In this paper, we focus on the dilemmas with some of the latest innovations in algorithmic surveillance, namely facial recognition technology and digital fingerprinting, which are now being introduced as part of the EES in the Schengen countries motivated by the need to increase security (Andrejevic & Zala, 2021; Mau, 2022; Trauttmansdorff, 2024). However, the technologies are also associated with concerns about far-reaching risks and threats to key democratic values such as the privacy of individuals (Ball & Webster, 2020; Eneman et al, 2022). Privacy is closely linked and intertwined with surveillance and, in line with previous research, we recognize that the term privacy in today's society is often defined as the right to control information about oneself (Bennett & Raab, 2020).

The technologies now being introduced with facial recognition and digital fingerprinting refer to automated systems for identifying human biometrics (Andrejevic & Zala, 2021, Smyth, 2019). Human facial images and fingerprints constitute biometric data that is unique and cannot be changed naturally.

Against this background, the aim of this paper was to explore ethical and sociotechnical dilemmas associated with the introduction of biometric technologies, such as facial recognition and digital fingerprints, at airport border control, which is also part of an emerging supranational infrastructure.

The empirical work consists of field visits, interviews and document analysis. The paper is structured as follows: The following section outlines the supranational border infrastructure taking shape within the Schengen area through the implementation of the Entry/Exit System. This is followed by a description of the research design. The subsequent section presents the empirical findings, and the chapter concludes with reflections of the key insights and their broader implications.

2 The Rise of a Supranational Infrastructure for Border Control

The concept of the border has been redefined and further developed in recent decades, from a fixed territorial demarcation to a dynamic, socio-technical system of inclusion, exclusion, and control (Mau, 2022; Trauttmansdorff, 2024). Scholars (Amoore, 2021; Amon, 2022; Mau, 2022) have questioned the traditional view of borders as static geopolitical lines, instead conceptualising them as relational and processual entities embedded within infrastructures of governance and algorithmic governance. Mau (2022) describes contemporary borders as filtering mechanisms, i.e. structures that sort individuals based on security policy risks rather than simply marking the limits of national sovereignty. He argues that airports function as critical nodes within these filtering systems. They are not merely points of entry or exit but are integral to the infrastructure that enables the sorting and surveillance of individuals. Through biometric data collection, automated screening, and real-time data analysis, airports exemplify the shift towards borders that are enacted through technopolitical infrastructures rather than solely through geographic. While Mau does not define airport borders explicitly, his analysis provides a framework to understand airports as pivotal spaces where contemporary border practices are operationalized, embodying the transition from static lines to dynamic processes of inclusion and exclusion.

Simmel (1997) argues that borders should not be understood primarily as a spatial construction with social consequences, but rather as a social distinction expressed spatially. This perspective corresponds well with Haggerty and Ericson's (2000) view of the surveillant assemblage, which conceptualizes surveillance as a process of abstraction, recombination, and restructuring of data. These assemblages are not confined to a single institutional actor or location, rather, they emerge through the convergence of technology, public and private actors, and policy and regulatory frameworks. The European Union's Entry/Exit System (EES), established under Regulation (EU) 2017/2226, is an illustrative example that materialises the ideas of such assemblage. Intended to control third-country nationals entering or exiting the Schengen Area, The EES will (once implemented) operate across national jurisdictions through data sharing and a biometric infrastructure that includes facial recognition, fingerprints and automated identity verification, among other things. We here describe this as a supranational infrastructure, a complex infrastructure consisting of systems that extends beyond individual member states and is sustained through interoperability, legal harmonisation, and continuous data sharing. Far from merely recording entries and exits, it also anticipates, evaluates, and classifies travellers in real time, redistributing the function of the border across databases, algorithms, and screening procedures. Such design reflects the distributed and deeply integrated nature of border systems under conditions of algorithmic control, where the border is no longer tied to a fixed geographical site but enacted through interconnected digital infrastructures (Amoore, 2021).

Amoore (2006) suggests the concept of the biometric border to underscore the dual nature of implementing this type of security policy within border control aimed at countering terrorism. Although the biometric border has become an almost ubiquitous feature in the fight against terrorism and crime, it also encompasses ambiguous and conflictual elements which makes it questionable. Through the concept of the deep border, Amoore (2021) adds another analytical dimension by placing border control within the logic of

algorithmic governance. Here, the border is no longer limited to a spatial threshold but operates deep within computational systems that govern mobility through predictions. In the EES, this form of governance occurs through biometric data, travel history, and embedded risk profiles. The border, in this view, becomes more ubiquitous, taking effect long before a traveller arrives at a checkpoint and persisting long after their departure.

Contemporary regimes of algorithmic border control as here with facial recognition and digital fingerprints mark a significant epistemic shift in how mobility and identity are governed. As Amoore (2020) argues, these systems do not seek to determine who someone is in any definitive sense, but rather whether an individual resembles a risk profile. This logic of approximation replaces certainty with probabilistic inference, whereby individuals are assessed not on the basis of concrete actions or identity, but on their computational likeness to previously defined markers of risk. In this sense, the border is no longer a stable geopolitical line but a dynamic threshold of algorithmic resemblance. Such a shift carries profound implications. First, it masks epistemic uncertainty behind a veneer of precision, producing the illusion of objectivity even as the outputs are shaped by opaque model assumptions and historical biases (Amoore, 2013; 2020). Second, it displaces responsibility by grounding decisions in future-oriented prediction rather than past actions or legal standards, thereby undermining traditional principles of accountability. Third, it exacerbates discriminatory outcomes, as approximation tends to replicate and amplify social inequalities embedded in the data and the classification systems themselves. As a result, the right to move becomes increasingly conditional, less a legal entitlement and more a function of not resembling a risk, thus transforming mobility from a right into a privilege granted by algorithmic discretion. In Amoore's terms, the algorithmic border does not observe; it infers.

Within the context of the European Union's Entry/Exit System (EES), the use of facial recognition and digital fingerprint as part of airport border control exemplifies how surveillance operates through an assemblage of technologies and institutions on different levels and logics. Following Haggerty and Ericson's concept of the surveillant assemblage, biometric border systems do not function as isolated tools but as part of an expanding network that fragments and reassembles personal data into categories (Mau, 2022). Here, operational images (Farocki, 2004) play a central role. These biometric data are not intended for visual inspection by humans, but are processed by automated systems that match identities, verify entitlements, and authorise or deny access in real time. They are images that act i.e. machine-readable traces embedded in an infrastructure of algorithmic governance (Amoore, 2024).

In this infrastructure, border control is no longer purely the domain of officials exercising discretionary judgement, but rather becomes a distributed and anticipatory process, in which decisions are delegated to interconnected systems that rely on biometric data to conduct the control (Mau, 2022). The operational image thus becomes a key element in a surveillant assemblage that blurs the boundaries between visibility and invisibility, authority and automation, state and private actors (Haggerty & Ericson 2000). This entanglement raises important questions concerning control, accountability, power and legitimacy in relation to the emerging surveillance practices as part of border control. Moreover, EU legislations have restricted the autonomy of national border control policies.

The EES is a large-scale EU border management system designed to digitally register non-EU-nationals who are crossing the external borders to the Schengen Area. It is based on biometric data collection such as fingerprints and facial images. It has been postponed several times in recent years, which has delayed its implementation, and it is now postponed to October 2025.

The aims of EES are to improve efficiency in border flows, reinforce security, and identify over-stayers. EES are targeting four main goals:

- Efficiency – Contributing to a modern border management with efficient flows at the external borders of the Schengen area. This will be done by automated border control gates and self-service kiosks.
- Security – Reinforce internal security in relation to terrorism and serious crime. The system will help to identify fraud. It will also assist in detecting travellers who pose a security or migration risk.
- Deal with the increasing number of travellers to the EU without increasing the number of border guards.
- Systematically identify individuals remaining in the Schengen area after their authorised stay (over-stayers). The system will automatically track the duration of stay.

The technical architecture of EES is based on:

- a central system that will operate a central database of biometric and alphanumeric data.
- a national uniform interface in each participating country.
- a secure communication channel between the central systems of the EES and of the VIS (Visa information System).
- a secure and encrypted communication infrastructure between the EES central system and the national uniform interfaces (identical interfaces for all Member States connect their border infrastructures to the EES central system).
- a data repository to obtain customisable reports and statistics.
- a web service to enable non-EU nationals to verify their remaining authorised stay.

3 Research Design

In this paper, we have investigated dilemmas in relation to the supra-institutional border infrastructure EES and the implementation of border technologies, such as facial recognition and digital fingerprints, as part of external border control at a large Swedish airport. The empirical material that forms the basis for this paper has been generated through field visits, interviews, and document analysis (Alvesson & Deetz, 2021).

Early in the empirical fieldwork, we realized that the implementation of the EES is surrounded by political sensitivity. To ensure confidentiality for the respondents included in the study, we do not name the airport but refer to it as a large Swedish airport with an external border. The research has also undergone ethical review and is approved by the Swedish Ethical Review Authority. In addition, the authors have extensive experience in conducting fieldwork involving sensitive data, and solid experience of empirical

fieldwork in relation to the implementation of new technologies within law enforcement agencies (Eneman & Ljungberg 2025). The research has been conducted in accordance with good research practice (Swedish Research Council, 2024; Hammersley & Traianou, 2012) in terms of information, confidentiality, voluntary participation and informed consent.

We have conducted four field visits, consisting of about 3 h each time, between Autumn 2022 to Spring 2025. The field visits took place at a major Swedish airport that is preparing for the implementation of biometric technologies i.e. facial recognition and digital fingerprints as part of border control. In the preparation for the formal implementation (currently planned for October 2025) the technologies have been tested as part of the border control. The specific airport being investigated constitutes an external border between the Schengen countries and so-called third countries (i.e. countries outside the EU) and constitutes an important traffic hub and a geographically delimited location where the border police are tasked with controlling passengers and the Swedish Customs are tasked with checking goods.

In addition to the field visits, we have conducted qualitative interviews with following actors:

- Border police officers (6 respondents) at the airport
- Border controllers (8 respondents) at the airport
- National coordinators within the Police Authority (4 respondents) responsible for the EES, who work nationally with the implementation of the EES with responsibility for coordination, technical and legal issues
- Supervisory authorities responsible for supervising the technology that is being introduced and will be used (2 respondents) and the policy level regarding legislation (1 respondent)
- Finally, we have interviewed people (2 respondents) within the Police Authority who are involved in the training of border controllers

During all interviews, we have taken field notes, and there were usually two researchers participating during the interviews. Immediately after the interviews, we went through our notes and supplemented them. Each interview took between 30–60 min.

We have also analysed official documents linked to the EES and the biometric technologies. Previous research (Bowen, 2009; Alvesson & Deetz, 2021) has highlighted that documents can be a rich source of data and emphasizes that it is important for researchers to read documents through a critical lens, i.e., not to view formal documents as “neutral” but to show awareness that the documents were created and conditioned for a specific purpose. The content of documents should be understood in their broader organizational, political, and social context, and the reader should reflect on whose voices are prioritized.

We have used thematic analysis (Braun & Clarke, 2021) to identify emerging patterns and themes linked to the dilemmas with the EES as a large complex supra-institutional border infrastructure and the implementation of the biometric technologies - facial recognition and digital fingerprints. We began by reading all the material (field notes and collected official documents); followed by us (the authors) discussing our interpretations of the material together. To formulate a more detailed understanding of what we considered

to be central and during that process we continued to identify relevant themes. These themes are presented in the next section.

4 Results

While Swedish border control operates in accordance with national laws and standards, it is nevertheless closely integrated with broader EU-level initiatives. Although technologies for airport border control have been identified, Sweden has yet to implement automated systems, in contrast to several other Schengen countries where components of the EES, such as automated border control, are already in use. The EES has been postponed several times, and is now supposed to launch in October 2025, which shows the complexity of organizing the implementation of such a complex supra-institutional infrastructure simultaneously in 29 countries. Key to EES and smart airport border control in general is biometric technologies such as fingerprints and facial recognition, together with a database to register all non-Schengen travellers. A number of issues related to the EES and biometric technologies emerged during the field visits and document analysis.

The Airport Border as a Material and Institutional Assemblage

Airports are complex configurations of physical arrangements, digital infrastructures and institutional settings. One important actor in Sweden is Swedavia, a Swedish state-owned company responsible for owning, operating, and developing Sweden's major airports. Swedavia is responsible for airport facilities operation and management, as well as procurement of basic technology and infrastructure, for example in the security control. They have the responsibility for flight security, as well as some of the airport security, which is delegated to a private security company while the police have the general responsibility for airport security. Swedavia owns the passport kiosks, but the police own the content in the kiosks. Both Swedavia and the police have their own cameras which they monitor and control. Thus, Swedavia is responsible for running the airport facilities and airport security infrastructure, while the border police is responsible for passport controls, immigration, and apprehending individuals with alerts, i.e. issues of entry and exit and related infrastructure. Then the customs are responsible for the flow of goods. How these actors can interact and share information is dependent on regulations, complexity in digital infrastructure, professional discretion, organisational responsibility and institutional powers.

Efficiency in Passenger Flows versus Efficiency in Security

The aim of automated border control is to be able to manage the increasing flow of non-Schengen passengers, without increasing the number of border guards. In Sweden there are no immediate plans to implement fully automated gates or self-service kiosks. The biometrical system will rather support the identification process in passport control and support the border police in their decisions. This leads to an increased security in terms of easier detection of identity fraud (e.g. look-a-like fraud). It also leads to an increased need of border guards, since each identified fraud requires extra work, taking the suspect to a second line room for further investigation. The successful identification of identity fraud is based on a combination of the physical arrangement of the passenger flow, where the "snake" form will give the border guards an overview of the que and detect suspicious behaviour. In these cases, the facial recognition system gives the border guards

a tool to easier detect look-a-like fraud. It is possible to manipulate the biometric facial image in the passport to fool the facial recognition algorithm, but this is still very rare, since it is a complex and very expensive task. With this use of the technology, security is increased, but not efficiency in flows and personnel. It also relies on a combination of technology and human judgement, as well as follow up activities of border guards.

Increased Documentation

The use of biometric technology is expected to lead to increased documentation at border controls, especially as so-called enhanced controls are expected to increase. Enhanced control means that the individual is accompanied to a separate room where two border control officers ask in-depth questions about identity and the purpose of entering Sweden. In addition, the travel documents are carefully examined with a microscope.

Technical and Infrastructure Challenges

These are linked to the establishment of a robust infrastructure and the integration of new biometric systems with existing infrastructure and to reach interoperability.

-They are also linked to specific technologies in specific user contexts. One example is that the placement of the physical camera used for testing facial recognition gives rise to challenges and extra manual work, as the height needs to be adjusted according to the individual's height in order to be able to read the face properly. Another example is problems with image quality in certain cameras.

Differentiation in Member State Readiness (and Willingness) to Implement EES

Different member states have different readiness and willingness to implement EES. For high traffic airports upgrades tend to be more complex to perform without delays. In Sweden delays could also be related to an expectation that better technology will arrive, and that therefore is better to wait. This relates to classical challenges in infrastructure development and interoperability, since digital infrastructure is intimately intertwined with organizational structures, procedures and local work practices, and thus shared among distributed actors with multiple and different needs and interests.

Legal and Regulatory Hurdles

Swedish law had to be adjusted to allow for the collection and processing of biometric data such as fingerprints and facial images. This means that the Police Authority will be responsible for providing access to the Entry/Exit System for other authorities and for handling the processing of personal data within the system. Furthermore, it is clarified how the regulations relate to other data protection legislation and under what circumstances a foreign national is required to provide fingerprints and be photographed in order to be checked against the EES.

Detection and Prevention of Serious Crimes and Terrorism

To prevent, detect or investigate terrorist offences or other serious criminal offences, designated law enforcement authorities and the European Union Agency for Law Enforcement Cooperation (Europol) may make a request to consult EES data. Access to the EES should not be prevented where a hit is obtained during the consultation of a relevant database prior to accessing the EES. "...a search of the database of the EES should be deemed proportionate if there is an overriding public security concern. Any search must be duly justified and proportionate in the light of the interest invoked." (Regulation (EU) 2017/2226, p. 23). When biometric data collected through the EES is processed

for law enforcement purposes, it falls under the scope of Directive (EU) 2016/680—commonly known as the Law Enforcement Directive (LED). This regulation governs the processing of personal data by competent authorities for criminal justice purposes, allowing for derogations from individual data rights such as access, rectification, and erasure. The LED introduces a layer of legal complexity in which state and supra-state actors operate with significant discretion, highlighting the tension between data subject rights and institutional mandates for control and security.

Privacy, Sorting and Profiling

Border control includes profiling, i.e. the detection of suspicious behavior during queuing or in conversations with border guards. This is also linked to risk assessment. Border control evaluates the risk for illegal migration and threats to the general order or security upon entry or exit. Also risks for human trafficking and non-sanctioned travels without the legal guardian's permission (child abductions) is taken into consideration when a minor's right to entry or exit is evaluated. Here border control collaborates with the social services and it is not allowed to discriminate against anyone based on gender, ethnic background, religious beliefs, disability, age or sexual orientation. Thus, the evaluation that an individual or group of individuals is to be selected for second line control, cannot be based on ethnicity alone.

Travelers' Privacy

It is not uncommon for travellers to express hesitation about being photographed at border control. This can lead to a discussion with travellers where border control officers must inform about the new regulations that require the collection of biometric data. Challenges regarding traveller privacy were identified in the document analysis. Thus, crossing border control can be perceived as a stressful situation for travellers, which may affect the travellers' ability to absorb the information provided, for example regarding the collection of biometrics. It also appears that this stress can contribute to uncertainty among third-country nationals who are registered at border control. The same document also pointed out the importance of respecting human dignity and not carrying out biometric collection at border control in a discriminatory manner.

Logistic Concerns Linked to Potential Disruption of the Airport as Commercial Arena

These concerns are related to disturbances in tourism, trade and the overall flow of people across borders.

5 Concluding Remarks

This paper has shown that the introduction of biometric technologies as part of border control, and within the broader framework of the EES, constitutes a profound shift in how borders are enacted, understood, and governed. Borders are no longer confined to fixed spatial demarcations, but now function as dynamic, algorithmically mediated infrastructures composed of institutions, databases, and predictive systems. In this transformation, facial recognition and digital fingerprinting should not be understood merely as tools for verifying identity, but as integral components of an expansive surveillant assemblage that classifies, filters, and governs individuals' mobility in advance of their physical movement.

Drawing on Amoore's (2020) concept of approximation, we argue that these systems operate not by establishing absolute identities, but by assessing the likeness of an

individual to predefined patterns of risk. Rather than offering epistemic certainty, biometric border systems infer probabilities, producing classifications based on resemblance to aggregated data profiles. This shift from identity to risk-based inference introduces a new epistemology of control, in which algorithmic outputs stand in for legal or evidentiary standards. As Amoore warns, such systems render borders into zones of computational judgement, where the right to cross is contingent not on who one *is*, but on whom one appears *similar* to.

While these technologies are often framed as neutral or even necessary solutions to enhance security, our findings highlight a series of ethical and sociotechnical dilemmas, for example tensions between efficiency, security, and the protection of civil liberties. The risk of normalising intensified surveillance at international borders, under the guise of innovation, demands sustained ethical scrutiny, especially given the opacity and predictive nature of the systems involved.

Finally, the deployment of biometric technologies raises legal and moral questions that go beyond compliance with data protection frameworks. Issues such as proportionality, non-discrimination, and the right to erasure, as codified in the General Data Protection Regulation (GDPR), become particularly salient when biometric identifiers such as facial images and fingerprints are collected and stored. We argue that airports no longer function merely as sites of control, but as technopolitical spaces, where public values, private interests, and institutional power intersect in complex and often opaque ways.

As biometric borders continue to evolve, there is a pressing need for an ongoing, critical interrogation of the normative assumptions, power asymmetries, and epistemic practices embedded in algorithmic border control (Amoore 2024). Without such reflection, we risk entrenching surveillance systems that not only monitor movement but fundamentally reshape how personhood, risk, and legitimacy are understood at the border.

If borders are no longer about who you are, but about who the system believes you might be, what does that mean for the future of rights, accountability, and democratic oversight?

Acknowledgments. We would like to thank *the research program Future Challenges in the Nordics – People, Culture and Society* is financed by the Society of Swedish Literature in Finland, The Finnish Cultural Foundation, The Swedish Cultural Foundation in Finland, Stiftelsen Brita Maria Renlunds minne, Riksbankens Jubileumsfond and The Kamprad Family Foundation for Entrepreneurship, Research & Charity and *the Swedish Research Council for Health, Work and Welfare (Forte)* for funding this research.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

Amoore, L., Campolo, A., Jacobsen, B., & Rella, L.: The deep border. world model: On the political logics of generative AI. *Politica* (2024)

- Andrejevic, M., Zala, V.: Seeing like a border: biometrics and the operational image. *Digit. Cult. Soc.* **7**(2), 139–158 (2021)
- Ball, K., Webster, W.: *Surveillance and Democracy in Europe*. Routledge (2020)
- Bauman & Lyon (2013)
- Bennett, C., Raab, C.: Revisiting the governance of privacy: contemporary policy instruments in global perspective. *Regul. Gov.* **14**(3), 447–464 (2020)
- Braun, V., & Clarke, V.: *Thematic Analysis : a Practical guide*. SAGE (2021)
- Donnan, H., Wilson, T.M.: *Borders: Frontiers of Identity, Nation and State*. Berg (1999)
- Eneman, M. & Ljungberg, J.: AI and governance dilemmas for law enforcement agencies. In Iveroth, E., Lindvall, J & Magnusson, J (eds.) *Leading Digital Transformation: Management, Governance and Control*. Routledge. (2025)
- Eneman, M., Ljungberg, J., Raviola, E., Rolandsson, B.: The sensitive nature of facial recognition: tensions between the Swedish police and regulatory authorities. *Inf. Polity* **27**(2), 219–232 (2022)
- Farocki, H.: Phantom images. *Public* **29**, 12–22 (2004)
- Guild, E.: Schengen borders and multiple national states of emergency: from refugees to terrorism to COVID-19. *Eur. J. Migr. Law* **23**(4), 385–404 (2021)
- Guild, E., Brouwer, E., Groenendijk, K., Carrera, S.: What is happening to the Schengen borders?, *Liberty and Security in Europe*, No. 86 (2016)
- Hammersley, M., Traianou, A.: *Etik i kvalitativ forskning*, SAGE Publications Ltd. (2012)
- Haggerty, K. D., & Ericson, R. V.: The surveillant assemblage. *Br. J. Sociol.* **51**(4), 605–622 (2000)
- Jeandesboz, J.: Border security and the politics of risk: The EU and the securitisation of irregular migration. In: Lazaridis, G., Campani, C., Benveniste, A. (eds.) *The Securitisation of Migration in the EU*, pp. 115–136. Palgrave Macmillan (2016)
- Khan, N., Efthymiou, M.: The use of biometric technology at airports: the case of customs and border protection (CBP). *Int. J. Inf. Manag. Data Insights* **1**(2) (2021)
- Lyon, D.: *Surveillance as Social Sorting: Privacy, Risk, and Digital Discrimination*. Routledge (2003)
- Lyon, D.: *Identifying Citizens: ID Cards as Surveillance*. Polity Press (2009)
- Mau, S.: *Sorting Machines*, John Wiley and Sons LTD. (2022)
- Simmel, G.: The sociology of space. In: Frisby, D., Featherstone, M. (eds.) *Simmel on culture: Selected writings*, pp. 137–170. Sage (1997)
- Smyth, S.: *Biometrics, Surveillance and the Law*, Routledge. (2019)
- Trauttmansdorff, P.: *The Digital Transformation of the European Border Regime: The Power and Perils of Imagining Future Borders*, Bristol University Press (2024)



Quizly: Transforming Quiz Experiences with Multi Modal Inputs for Differently Abled Users

Farzana Jabeen^(✉), Sidra Sultana, Tahira Anwar Lashari, and Mehvish Rashid

National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
farzana.jabeen@seecs.edu.pk

Abstract. This research presents Quizly, a multimodal Android quiz application designed to enhance educational accessibility, particularly for differently abled users. The objective is to improve user interaction in quiz-based learning through the integration of machine learning and multimodal input methods. Quizly incorporates hand gesture recognition, voice navigation, and gaze tracking, leveraging Convolutional Neural Networks (CNNs) and a cross-platform tech stack including Python, React Native, MongoDB, Express, and Node.js. The system was developed using an incremental and iterative methodology, with usability testing conducted among 30 participants. Results indicate high user satisfaction and effective performance across input modes. The app achieved an overall input recognition accuracy of 89% across all participants. The findings highlight the value of user-centric design and responsive interaction in educational applications, offering a foundation for future innovations in inclusive learning technologies.

Keywords: Multi-modal interaction · Differently abled users · Universal Design for Learning · Inclusive Technology · Mobile application Development

1 Introduction

The world now has a very strong global movement of people who stand for the education of the computer science basics for the students at a very young age, along with the foundation skills like reading, writing and mathematics. This overarching movement has been given much attention to by the international community, stressing the importance of reforming current computer science teaching methods [1]. The main goal is to establish a platform which will allow computer science education to be provided at every level of the education system, with the possibility of access to all students. This validation rather highlights a more purposeful direction towards the conforming of educational activities with the changing requirements of the digital age. In the rapid evolving landscape of educational technology, Quizly proves to be an example of the innovation and

inclusiveness. It is the time being that technology is changing learning environment dramatically [2]. Therefore, it is necessary to cater for the different needs of students including those with visual impairments. Quizly, which is a cutting-edge Multimodal Android Quiz Application, stands apart from the rest by delivering a convenient and inclusive learning platform to visually impaired students. The creation of Quizly is based on a detailed research methodology and user-friendly and cognitive support first.

This article presents the research framework behind the development of Quizly, a platform designed for students with visual impairments and blindness. It highlights UI/UX design elements and the integration of gaze tracking and gesture recognition to boost user engagement. Quizly demonstrates a seamless blend of design, technology, and user-focused features, offering real-time interaction, cross-platform support, and intuitive interfaces. Hosting on a cloud platform ensures scalability, reliability, and security. The article also details the gaze tracking model and its training process, enabling users to accurately select answers. Hand gesture input further enhances interactivity, making quiz-taking engaging and intuitive. Finally, the results showcase improvements in gaze and gesture recognition accuracy during training. The research paper makes significant contributions in the following areas:

- Multi-modal Interaction: Seamlessly integrates gaze tracking and hand gesture recognition, providing an inclusive and captivating quiz experience.
- User-Centric Design: Introduces the idea of user-centered design through Figma, allowing more users with different abilities to access the interface.
- Innovative Model Training: Utilizes cutting-edge machine learning techniques, with a particular focus on CNNs, to ensure precise gaze tracking and gesture recognition. These advancements contribute to an enhanced educational impact within the application.

Structure of the Research – The research is organized into distinct sections: The introduction is given in Sect. 1. Literature review is provided in Sect. 2. Section 3 presented the research methodology of this study. Implementation, User interface and Gaze tracking and Model Training is given in Sect. 4 and 4.3, respectively. Results are provided in Section User Study Evaluation and finally, the paper is concluded in Sect. 6.

2 Literature Review

Assistive technologies and Universal Design for Learning (UDL) are closely related concepts with some overlap [3]. Assistive technologies enable students with disabilities to access content through alternative means, such as tools that support those with sensory or auditory impairments. The implementation of UDL in higher education aims to broaden curriculum accessibility to a diverse audience. Ramirez et al. [4] emphasize providing such technologies across higher education institutions to support inclusive learning. Jenkins [5] argues that UDL can be fully effective only when all learners receive equitable access to

knowledge, even when traditional teaching methods are disrupted. Gamification has gained prominence in education for its ability to enhance learning by boosting motivation, engagement, participation, and interactivity [6–8]. When implemented effectively, it increases intrinsic motivation and supports educators at all levels [9–11]. Numerous studies explore its educational applications and benefits [12–15], particularly its positive impact on student outcomes [16, 17]. Addressing low motivation in traditional classrooms has become essential [18, 19], as conventional methods often fail to foster meaningful engagement [20–22]. Therefore, integrating gamification is increasingly recommended as a viable solution. Gamification enhances learning by introducing fun through challenges, rewards, and friendly competition, which fosters active engagement [23]. It also helps learners develop critical thinking and multitasking abilities [24]. Moreover, gamification provides valuable data on student progress, aiding educators and stakeholders with accurate and timely insights [25]. Features like scores, badges, and rankings deliver immediate feedback, encouraging participation and task completion [26]. Educators can track learning outcomes and offer formative feedback [27, 28]. Studies such as [29] propose inclusive design frameworks, while others like [30] and [31] explore ML-based and hybrid models to boost engagement. This approach supports formative assessment by highlighting both learning patterns and game design effectiveness. A summary is presented in Table 1.

Table 1. Comparison of Related Work in Quiz Application.

Authors	Quiz	Gaze	Hand	Supported	Machine	Universal	Gamification
F. Fovet [3]	√	×	×	Universal Design Learning	×	√	√
E. Lee [8]	√	×	×	Design Framework	×	√	√
DR. Sanchez [12]	√	×	×	Online Gamification Testing	×	×	√
S. Bai [17]	×	×	×	Gamification Meta-Analysis	×	×	√
AF. Azizah [29]	√	√	×	Design Framework	√	×	√
K. Duggal [30]	√	×	×	ML based Application	√	×	√
KS. Narayanan [31]	×	×	×	Hybrid Gamification	√	×	√
Proposed	√	√	√	Multi-Modal Quizly	√	√	√

3 Research Methodology

3.1 App Development

Developing learning apps for differently abled users requires a thoughtful, inclusive methodology that ensures both functionality and accessibility. Quizly’s screen design follows a user-centric layout, with each interface divided into six equal zones to simplify navigation for visually impaired users. This consistent

structure minimizes confusion and enhances muscle memory, making interaction intuitive and efficient. In Quizly, design elements like screen division, text-to-speech, high-contrast visuals, large fonts, and haptic feedback were implemented to support visually impaired users. These features reduce cognitive load and empower users to navigate independently. Ethically, this approach reflects a commitment to equitable access and user autonomy, embedding inclusivity not as an add-on but as a core design principle. A diverse development team and careful prototyping ensured the app serves as a genuinely accessible educational tool. In addition to its ethical design, Quizly incorporates robust technical features to enhance accessibility. The app supports multi-modal interaction through gaze tracking, hand gesture recognition, and voice commands, enabling users with various impairments to interact without relying on traditional touch input. Its text-to-speech engine allows visually impaired learners to consume content audibly, while customizable settings let users adjust voice speed, contrast, and interface elements to suit their individual needs. These capabilities are seamlessly integrated using a cross-platform architecture built with React Native, Python-based machine learning models, and a scalable Node.js backend—ensuring both performance and inclusivity at scale, the architecture of the system is illustrated in Fig. 1.

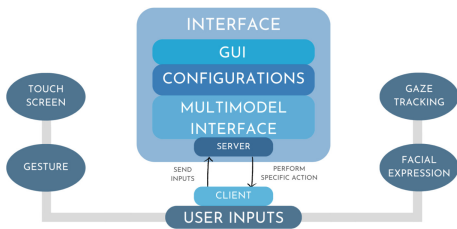


Fig. 1. Architecture of the system

3.2 Integrated Design and Development Approach

The development of the Multimodal Android Quiz Application followed a unified approach, covering UI/UX design, front-end and back-end development, database management, and machine learning integration. The UI/UX design was created with Figma, focusing on aesthetics and user engagement, incorporating hand gesture recognition, gaze tracking, and quiz elements. Iterative feedback improved the overall experience. React Native was used for front-end development to enable cross-platform support on Android and iOS, providing seamless integration with machine learning models. On the back-end, Node.js with Express delivered a scalable infrastructure to handle user interactions, data flow, and machine learning communication. MongoDB was chosen for its flexible,

document-based structure that supports scalability and efficient data management. Machine learning models for hand gesture and gaze-based quiz answer recognition were developed in Python, selected for their effectiveness in enabling accurate and interactive application features.

3.3 User Design Centric

The Multimodal Android Quiz Application was developed with a strong emphasis on user-centered design. Key strategies included:

- **User Research and Personas:** Comprehensive user research, including surveys and interviews, guided the creation of personas representing diverse user groups. These personas influenced design decisions to meet varied needs and preferences.
- **Iterative Prototyping:** Prototypes were developed using Figma, allowing for early visualization and iterative refinement based on user feedback. This approach ensured the design was both functional and visually appealing.
- **Accessibility Considerations:** Features such as voice commands and adaptable text sizes were incorporated to enhance accessibility for users with disabilities, ensuring compliance with accessibility standards.
- **Usability Testing:** Usability testing was conducted at various stages to evaluate how easily users could navigate the application and use its features. Feedback from these tests was used to refine and improve the user experience.
- **User Feedback Integration:** Ongoing user feedback was analyzed and used to make iterative design improvements, ensuring the final product was well-aligned with user needs and expectations. The application's user interface was crafted in Figma, considering feedback from user acceptance testing to create an intuitive and engaging experience.

3.4 Integration and Implementation

The development of the Multimodal Android Quiz Application blends design, technology, and user-focused features. React Native was selected for front-end development, allowing smooth API-based integration with machine learning models. This supports intuitive interactions, including gesture and gaze tracking. Hand gesture recognition, powered by CNNs, enables accurate quiz navigation, while gaze tracking allows users to select answers with eye movements. React Native ensures cross-platform compatibility across Android devices, enhancing accessibility. Real-time feedback is achieved via API integration, creating an interactive learning experience. The app is deployed on a cloud platform, ensuring scalability, reliability, and secure service with ongoing updates.



Fig. 2. Screens from the Application: Sign-Up, Login, and Dashboard

4 User Interface of the Application

4.1 Screens from the Application: Sign-Up, Login, and Dashboard

Users can create an account by providing a username, email address, and password on the signup screen, illustrated in Fig. 2(a). As an alternative, users have the option to conveniently register with their Facebook or Google accounts.

Users can quickly log in by entering their account email address and password on the login screen, as seen in the Fig. 2(b). They can also use their Facebook or Google accounts to log in as an alternative.

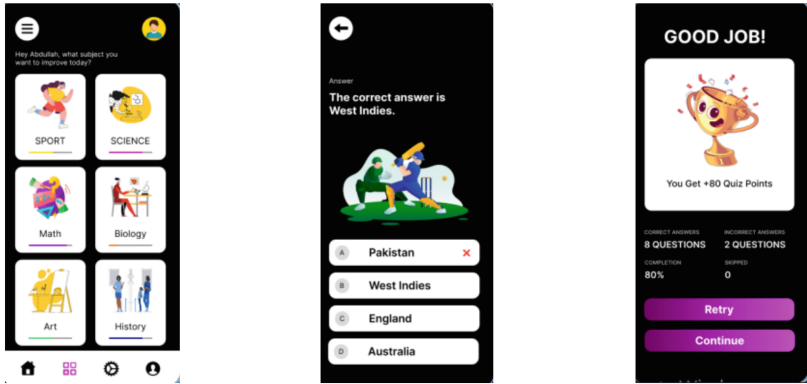
The dashboard interface, as shown in Fig. 2(b), greets users with their name and a cheerful message, followed by an inviting “Let’s Play” prompt. It displays the user’s rank, total points, and current level, offering a quick overview of progress. Recent quizzes are listed for easy review, and a prominent ‘Continue’ button encourages users to resume their quiz journey. The layout ensures easy access to key information while maintaining an engaging experience.

4.2 Quiz Interface: Selection, Question and Result Screens

Figure 3(a) illustrates a visual interface showcasing various quiz categories such as Sports, Science, Mathematics, Biology, Art, and History. Students select a subject based on their interests to proceed with the quiz.

After selection, the quiz screen appears as shown in Fig. 3(b), displaying a question with multiple options—e.g., Pakistan, West Indies, England, and Australia—related to a sports topic. Feedback is also provided, indicating whether the answer is correct; in this example, the correct answer is West Indies.

Upon completion, the Quiz Results Screen (Fig. 3(c)) displays a message like “Good Job!” with earned points (e.g., “+80 Points”) and a breakdown of correct



(a) Quiz Selection (b) Quiz Question Screen (c) Quiz Result Screen

Fig. 3. Quiz Interface: Selection, Question and Result Screens

and incorrect answers (e.g., 8 correct, 2 incorrect). Two buttons—“Retry” and “Continue”—allow users to retake the quiz or return to the home screen.

4.3 Multi-modal Training

Gaze Tracking. The Gaze Tracking for Answer Selection feature uses advanced technology to improve the quiz experience by tracking eye movements, allowing precise answer control. The model is trained on the ARGaze dataset containing 80,000 images (32×32 pixels), split 80% for training and 20% for testing. Its architecture combines CNN with facial landmark detection, pupil tracking, and calibration to enhance accuracy. Training uses rms-prop optimizer and categorical crossentropy loss, suitable for multiclass classification. Achieving 83% accuracy, the model reliably interprets gaze for precise quiz answer selection.

Hand Gestures. The Quizly app offers an engaging quiz experience through Gesture Recognition using computer vision algorithms. This feature identifies hand gestures, enabling intuitive interaction for selecting answers and accessing features. The system is trained on a dataset of 25,000 videos and 100,000 images (144×144 pixels), providing strong input for recognizing diverse gestures. Gesture use adds interactivity, creating a dynamic quiz environment that boosts user participation and enjoyment.

Model Training. Quizly leverages CNNs within its machine learning framework to enhance predictive accuracy, question analysis, and overall quiz performance. CNNs enable the system to recognize complex patterns in quiz data, improving answer prediction precision and enriching the user experience. This advanced use of CNNs goes beyond traditional methods, creating a dynamic and accurate quiz environment. Quizly’s integration of sophisticated model training

highlights its commitment to delivering a high-quality and engaging quiz experience.

Training Results. The outcomes of the Gaze Tracking and Hand Gestures models are represented in Figs. 4(a) and 4(b), respectively. An observation is that in both Gaze Tracking and Hand Gestures, the initial stages of training exhibit considerably high values for both training and validation loss. However, as the training iterations progress, a discernible trend emerges, depicting a gradual reduction in the loss values. This signifies the models’ increasing proficiency in minimizing errors and optimizing their performance over time.

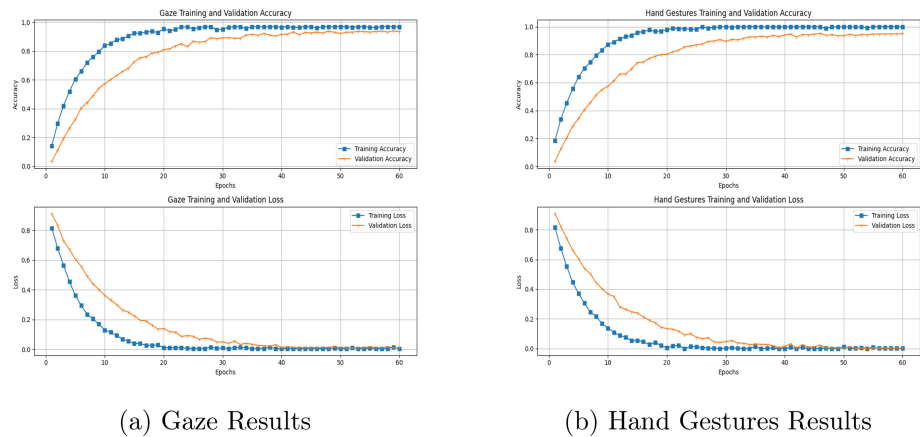


Fig. 4. Results of Gaze and Hand Gesture Recognition

5 User Testing and Evaluation

5.1 User Acceptance Testing

UAT Plan. User Acceptance Testing (UAT) was conducted to validate the Multimodal Android Quiz Application’s readiness for release and to gather feedback from end users. The UAT plan outlined the objectives, scope, and approach for testing. The key steps included:

- **Test Scenario Identification:** Test scenarios were defined based on core features such as hand gesture recognition, gaze tracking, quiz functionality, and UI responsiveness.
- **Test Data Preparation:** Representative test data were prepared, including sample gesture images, gaze patterns, and quiz sets, to simulate realistic user behavior.

- **Participant Recruitment:** Participants were recruited using purposive sampling to ensure diversity across user abilities. This included individuals with visual impairments, motor impairments, and individuals without disabilities.
- **Task Assignment:** Participants were assigned specific tasks: logging in, creating a quiz, interacting via gesture/gaze input, submitting a quiz, and reviewing scores. These tasks were designed to test both accessibility and functional aspects.
- **Test Execution:** A group of representative users executed the test cases, using all modes of interaction supported by the app.
- **Feedback Collection:** A mixed-methods approach was employed. Quantitative usability data were collected using a 10-point Likert-scale questionnaire based on the System Usability Scale (SUS), while qualitative feedback was gathered through semi-structured interviews.
- **Monitoring and Reporting:** All sessions were monitored, and issues were documented with severity levels and user feedback. A detailed report summarized findings and proposed refinements.

5.2 User Study Evaluation

To evaluate usability and accessibility, a comprehensive user study was conducted involving 30 volunteer participants, categorized as follows:

Participants

- 10 with visual impairments
- 10 with motor impairments
- 10 without disabilities (control group)

Methodology. The study included the following stages:

1. **Introduction and Training:** A 15-minute guided tutorial introduced the app and demonstrated interaction modes.
2. **Assigned Tasks:** Each participant performed five core tasks: user login, taking a quiz using at least two interaction modes, submitting results, reviewing scores, and adjusting app settings.
3. **Usability Measurement:** Participants completed a 10-question SUS-based Likert-scale questionnaire, scoring usability from 1 (strongly disagree) to 10 (strongly agree). Additional comments were collected via interviews.

Results

- **Accuracy:** The app achieved an overall input recognition accuracy of 89% across all participants.
- **Engagement:** Users reported high engagement due to multimodal interaction, particularly gaze tracking and gesture input, enhancing inclusivity.
- **Usability:** Positive feedback highlighted the intuitive layout and minimal learning curve.

ANOVA Test Justification: ANOVA (Analysis of Variance) was chosen to test if mean usability scores significantly differed across the three participant groups. This test is suitable for comparing more than two groups on a continuous dependent variable (Tables 2 and 3).

Assumptions Testing:

- *Normality:* Verified using Shapiro-Wilk test for each group ($p > 0.05$).
- *Homogeneity of variances:* Tested using Levene’s test ($p > 0.05$).

Usability Scores (Scale: 1–10):

ANOVA Results:

Interpretation:

Table 2. Usability Scores Across Participant Groups

Participant	Visual Impairments	Motor Impairments	Control Group
1	8	7	9
2	7	8	9
3	8	7	8
4	9	8	10
5	7	10	8
6	8	7	9
7	7	6	8
8	8	7	9
9	10	8	9
10	8	7	10

Table 3. ANOVA Test Results on Usability Scores

Source	SS	df	MS	F	P-value	F crit
Between Groups	2.333	2	1.167	4.500	0.041	3.885
Within Groups	14.833	27	0.549	—	—	—
Total	17.167	29	—	—	—	—

- F-value (4.500) > F crit (3.885)
- P-value (0.041) < 0.05

Conclusion: Since the null hypothesis is rejected, we conclude that there is a statistically significant difference in usability scores between participant groups.

6 Conclusion

The Android multimodal quiz application marks a significant advancement in educational technology. By integrating machine learning, gesture recognition, and gaze tracking, it transforms traditional quiz-taking into an interactive and engaging experience. Grounded in user-centric design and accessibility, particularly for visually impaired users, the application promotes an inclusive learning environment. The development process covering app design, model training, and user testing with 30 participants demonstrated strong usability and performance, validating the system’s effectiveness. This solid empirical foundation supports ongoing innovation. Future development should focus on enhancing CNN-based models to improve answer prediction and question analysis. Expanding dataset diversity and testing across broader user groups will further ensure inclusivity. Additionally, incorporating voice recognition and augmented reality can introduce new interaction modalities. Finally, personalized learning paths, collaborative features, and alignment with global accessibility standards are key to advancing this platform. These steps will help shape the next generation of inclusive and intelligent educational tools.

References

1. Prensky, M.: Education to better their world: Unleashing the power of 21st-century kids. Teachers College Press (2016)
2. Burbules, N.C., Fan, G., Repp, P.: Five trends of education and technology in a sustainable future. *Geography Sustain.* **1**(2), 93–97 (2020)
3. Fovet, F.: Universal design for learning as a tool for inclusion in the higher education classroom: tips for the next decade of implementation. *Educ. J.* **9**(6), 163–172 (2020)
4. Ramirez, J.: Educating individuals with intellectual disabilities: The us constitution, citizenship and disability (2021)

5. Jenkins, M.: What is UDL? In: *Unlocking Learning Potential With Universal Design in Online Learning Environments*. IGI Global, pp. 1–12 (2024)
6. Metwally, A.H.S., Nacke, L.E., Chang, M., Wang, Y., Yousef, A.M.F.: Revealing the hotspots of educational gamification: an umbrella review. *Int. J. Educ. Res.* **109**, 101832 (2021)
7. Parsons, J., Taylor, L.: Improving student engagement. *Current Issues in Education*, vol. 14, no. 1 (2011)
8. Lee, E., Hannafin, M.J.: A design framework for enhancing engagement in student-centered learning: own it, learn it, and share it. *Educ. Tech. Res. Dev.* **64**, 707–734 (2016)
9. Alsawaier, R.S.: The effect of gamification on motivation and engagement. *Int. J. Inf. Learn. Technol.* **35**(1), 56–79 (2018)
10. Manzano-León, A., et al.: Between level up and game over: a systematic literature review of gamification in education. *Sustainability* **13**(4), 2247 (2021)
11. Jones, M., Blanton, J.E., Williams, R.E.: Science to practice: does gamification enhance intrinsic motivation? *Act. Learn. High. Educ.* **24**(3), 273–289 (2023)
12. Sanchez, E., van Oostendorp, H., Fijnheer, J.D., Lavoué, E.: Gamification. In: *Encyclopedia of Education and Information Technologies*. Springer, pp. 816–827 (2020)
13. Luo, Z.: Gamification for educational purposes: what are the factors contributing to varied effectiveness? *Educ. Inf. Technol.* **27**(1), 891–915 (2022)
14. Alahmari, M., et al.: Trends and gaps in empirical research on gamification in science education: a systematic review of the literature. *Contemp. Educ. Technol.* **15**(3) (2023)
15. Oliveira, W., et al.: Tailored gamification in education: a literature review and future agenda. *Educ. Inf. Technol.* **28**(1), 373–406 (2023)
16. Huang, R., et al.: The impact of gamification in educational settings on student learning outcomes: a meta-analysis. *Educ. Tech. Res. Dev.* **68**, 1875–1901 (2020)
17. Bai, S., Hew, K.F., Huang, B.: Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educ. Res. Rev.* **30**, 100322 (2020)
18. Appleton, J.J., Christenson, S.L., Furlong, M.J.: Student engagement with school: critical conceptual and methodological issues of the construct. *Psychol. Sch.* **45**(5), 369–386 (2008)
19. Barkley, E.F., Major, C.H.: *Student engagement techniques: A handbook for college faculty*. John Wiley & Sons (2020)
20. Felder, R.M., Brent, R., et al.: Effective strategies for cooperative learning. *J. Cooperation Collab. Coll. Teach.* **10**(2), 69–75 (2001)
21. Rohrer, D., Pashler, H.: Recent research on human learning challenges conventional instructional strategies. *Educ. Res.* **39**(5), 406–412 (2010)
22. Shernoff, D.J.: *Optimal learning environments to promote student engagement* (2013)
23. Saleem, A.N., Noori, N.M., Ozdamli, F.: Gamification applications in e-learning: a literature review. *Technol. Knowl. Learn.* **27**(1), 139–159 (2022)
24. Aibar-Almazán, A., et al.: Gamification in the classroom: Kahoot! As a tool for university teaching innovation. *Front. Psychol.* **15**, 1370084 (2024)
25. Chan, G.L., Santally, M.I., Whitehead, J.: Gamification as technology enabler in SEN and DHH education. *Educ. Inf. Technol.* **27**(7), 9031–9064 (2022)
26. Arufe Giráldez, V., Sanmiguel-Rodríguez, A., Ramos Álvarez, O., Navarro-Patón, R.: Can gamification influence the academic performance of students? *Sustainability* **14**(9), 5115 (2022)

27. Zainuddin, Z., Shujahat, M., Haruna, H., Chu, S.K.W.: The role of gamified e-quizzes on student learning and engagement: an interactive gamification solution for a formative assessment system. *Comput. Educ.* **145**, 103729 (2020)
28. Moreira, H., Freire, M.L.L.: Promoting formative assessment with Quizizz: a classroom action research study. *Ciencia Latina Revista Científica Multidisciplinar* **8**(2), 590–604 (2024)
29. Azizah, A.F., Djunaidy, A., Siahaan, D., Suhariadi, F.: Improving the interaction of autistic children through eye tracking using gamification design framework. In: 2021 International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE), pp. 57–62. IEEE (2021)
30. Duggal, K., Gupta, L.R., Singh, P.: Gamification and machine learning inspired approach for classroom engagement and learning. *Math. Probl. Eng.* **2021**(1), 9922775 (2021)
31. Narayanan, K.S., Kumaravel, A.: Hybrid gamification and AI tutoring framework using machine learning and adaptive neuro-fuzzy inference system. *J. Adv. Res. Appl. Sci. Eng. Technol.* **42**(2), 221–233 (2024)

Security: Re-evaluating Best Practices of Cybersecurity



Comprehensive Approaches to Personal Data Protection Amid Evolving Cyber Threats

Sabina Szymoniak^(✉) and Mariusz Kubanek

Department of Computer Science, Czestochowa University of Technology,
Dabrowskiego 69, Czestochowa, Poland
`{sabina.szymoniak,mariusz.kubanek}@icis.pcz.pl`

Abstract. The protection of personal data is more critical in today's online security environment, given the rise in cyber threats and the extensive exploitation of user data. Intelligent cyberattacks, poor data management, and the growing application of Big Data-based profiling methods increasingly challenge organisations. This study provides a comprehensive overview of the development of personal data protection, encompassing legislation, technology, and organisational practices aimed at mitigating security risks. This study is a valuable addition, as it adopts a holistic approach. It integrates regulations such as the GDPR and the NIS2 Directive with advanced security controls. Such controls include encryption, anonymisation, and artificial intelligence-based threat detection. In addition, this paper explores the future applications of blockchain and quantum computing, examining how they can influence data protection strategies in the years to come. This study identifies critical gaps in current data protection strategies through an in-depth literature review and real-world case studies. It offers a blueprint for enhancing security controls. Our findings underscore the necessity for a comprehensive strategy that integrates adherence to the law, robust security measures, and user education to effectively protect personal data in an era of rapid technological advancement.

Keywords: Personal data protection · Encryption · Anonymisation · Cybersecurity

1 Introduction

Personal data has become one of the most treasured possessions in the digital realm. It impacts business models, personal relationships, and economic strategy. Organisations worldwide now collect and analyse tons of personal information, resulting in risk exposure as severe as hacking, identity theft, or mass surveillance. Thus, the real challenge in cybersecurity is strong personal data protection using legal, technological, and organizational measures [12].

Increasing sophistication in cyberattacks is among the most pressing cybersecurity threats. Modern cybercriminals employ advanced techniques to deploy

phishing [31], ransomware [17], and DDoS attacks [15] to obtain sensitive information. In addition, improper data management, such as storing information in an unencrypted form under outdated security protocols, further exposes this information to additional vulnerabilities. Additionally, the use of Big Data and Artificial Intelligence (AI)-powered profiling has heightened privacy concerns, including risks of discrimination and manipulation [5,23]. Figure 1 shows the key threats to personal data.

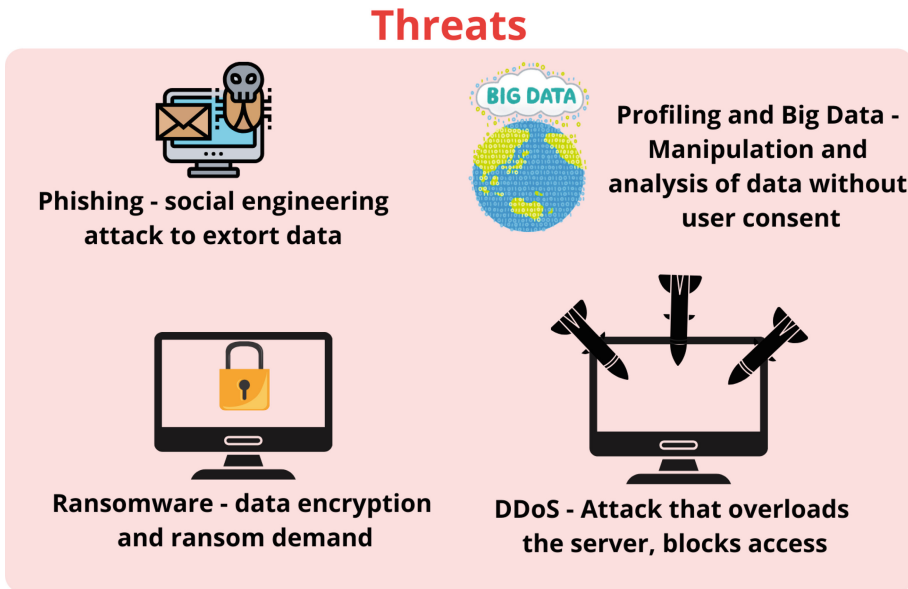


Fig. 1. Key threats to personal data.

Governments and regulatory bodies have introduced strict data protection laws to address these challenges. The General Data Protection Regulation (GDPR) [6] and the NIS2 Directive [24] establish high security standards, including mandatory data encryption, breach notification, and privacy-by-design principles. However, compliance alone is insufficient—technological advancements such as AI-based anomaly detection, blockchain for secure transactions, and quantum-safe cryptography are essential components of a comprehensive data protection strategy [10,26,28,30].

This paper systematically analyses personal data protection amid evolving cyber threats. The key contributions of this study are:

- a structured review of major cyber threats to personal data, including real-world cases.
- an integrated approach that combines legal compliance, security technologies, and user education.

- an exploration of emerging technologies such as AI, blockchain, and quantum computing in data security.
- identification of research gaps and future challenges in personal data protection.

The rest of this paper is organised as follows: Sect. 2 outlines key threats to personal data, whereas Sect. 3 shows protection mechanisms. Section 4 discusses a comprehensive approach to data security, while Sect. 5 addresses future challenges. Finally, conclusions will be drawn in Sect. 6 along with the recommendations for future work.

2 Main Threats to Personal Data

Threats to personal data are a huge problem in the modern world. As more data is processed and stored online in various databases, it can fall prey to cybercriminals and be used by them. We can highlight four main threat groups. The first group are cyberattacks. The most dangerous ones include:

- Phishing – fraudsters impersonate trusted institutions (for example, banks, courier companies, social networking sites) and send fake emails or text messages asking for personal data, logins or passwords. Clicking on a link leading to a fake site may result in the theft of login details and access to the victim's account [2].
- Credential Stuffing – an attack that involves using previously stolen logins and passwords to gain access to other user accounts. Many people use the exact login details across different services, which makes it easier for cybercriminals to gain access to email, banking, or social media accounts [3].
- Ransomware – malware that encrypts files on the victim's computer and demands a ransom to recover them. In an attack on a company, sensitive customer data may be leaked, and the organisation faces a dilemma – pay the ransom or risk disclosing the data [7].
- SIM Swapping – an attack that involves taking over the victim's phone number by making a duplicate SIM card. Criminals, posing as the owner of the number, contact the operator and convince them to issue a new card. In this way, they can take control of text messages, including codes for authorising bank transactions [11].
- Data Scraping – an unauthorised collection of personal data from public sources such as social media, internet forums or company websites. The collected information can be used to create detailed user profiles sold or used for personalised attacks, such as spear phishing [8].
- Malvertising – malicious advertisements placed on legitimate websites, which, when clicked, can redirect the user to a phishing site or automatically install malware on their device. In some cases, attackers exploit browser vulnerabilities to gain control of the system and access saved passwords or files containing personal data [14].

The second group of threats is related to improper data processing. Here, we can point out the following problems:

- breaches in organizations - resulting from human errors, such as sending data to the wrong recipients.
- abuse by employees - for example, unauthorized access to customer data.

We can also highlight threats arising from the development of technology that directly impact personal data protection. The two leading technologies are broadly used in the modern world.

- Internet of Things (IoT) - An increasing number of devices connected to the network, such as smart cameras, speakers, thermostats, or medical equipment, often lack adequate security. Manufacturers focus on functionality, and security issues are relegated to the background. Lack of updates, weak default passwords, or open network ports can lead to unauthorised access and leakage of sensitive user information [27].
- Artificial Intelligence (AI) - AI systems are trained on massive data sets, which may contain information about users. In some cases, models learning from real data may inadvertently store and reveal fragments of personal information. Additionally, algorithms used in the analysis of user behaviour can violate privacy, identifying people based on patterns of online activity. Another problem is the lack of transparency – users often do not know what data is being collected and how it is used [28].

The last group consists of threats related to legal and organisational data protection. Personal data may not fully comply with regulations without implementing GDPR requirements [6], such as the obligation to maintain a register of processing activities and user awareness, i.e., users' lack of knowledge about what data is collected and how it is used.

3 Personal Data Protection Mechanisms

There are many personal data protection mechanisms. We can divide them into the following groups of mechanisms.

Encryption and Secure Communication. Encryption is one of the basic mechanisms of data protection. Thanks to it, even if someone intercepts the communication, they cannot read its content. A popular example is end-to-end encryption (E2EE), used by messengers such as Signal or WhatsApp - only the sender and recipient can see the message. Self-protecting websites utilise TLS/SSL, which secures data communications between users and servers against unauthorised interception. Disk-level data is also encrypted. With tools like BitLocker and VeraCrypt, device data remains safe in the event of equipment theft [9].

Access Control and Authentication. Access control is a key element of data protection - not everyone should have access to everything. Many systems use

the so-called multi-factor authentication (MFA), which, in addition to a password, requires, for example, an SMS code or confirmation in the application (like Google Authenticator). Role-based access control (RBAC) is also utilised in companies, where users are assigned specific permissions based on their designated roles. We are also increasingly using biometrics, such as fingerprints or facial scanning, which make logging in easier while also increasing security [22].

Data Anonymisation and Pseudonymization. It is not always necessary to process complete personal data - often, anonymisation or pseudonymization is enough. In the first case, the data is transformed so that it can no longer be associated with a specific person (for example, names, Polish PESEL numbers, etc., are removed). In the second, instead of real information, random identifiers are used, but reversing them (for example, using a key) is still possible. Such techniques are used in medical research, statistical analysis or marketing to protect user privacy [19].

Legal and Regulatory Compliance. Personal data protection is not only a matter of technology but also of legal regulations. In the European Union, the GDPR plays a key role, which requires companies to process data by minimization, provide users with control over their information, and inform them about leaks. In the USA, CCPA introduced similar regulations. The Privacy by Design and Privacy by Default approach, i.e. designing systems with privacy protection in mind from the beginning, is also increasingly used [25].

Intrusion Detection and Prevention Systems (IDPS). Old methods like firewalls are not always enough. Modern defence systems that monitor what happens in real-time, picking up anything suspicious before any damage can be done, are needed. IDS (Intrusion Detection System) technologies analyse network traffic for potential threats, like unexpected logins or attempted attacks and pass these results on to administrators. Meanwhile, the IPS (Intrusion Prevention System) monitors network packets for unexpected behaviour by simply dropping them. The latest tactics increasingly utilise artificial intelligence to identify anomalies. In the present instance, this means finding better ways to protect systems against all sorts of new and unknown invasion strategies [28].

User Awareness and Training. The best security measures will not help if users make basic mistakes, such as using the same password everywhere or clicking suspicious links. That is why education is so important – both in companies and at the individual level. Training should cover phishing, safe network use, and data protection in social media. Regularly checking privacy settings and using password managers, which help create and store strong, unique logins, is also a good habit [21].

4 Comprehensive Approaches to Personal Data Protection

The issue of personal data protection can be considered in several aspects related to its processing, transfer, and the security mechanisms used. In the context of

legal regulations, everything depends on the specific country. To date, various approaches to personal data protection have been proposed.

Mestari et al. in [4] analysed privacy protection in machine learning systems and described the various threats associated with data processing, from attacks on models that may reveal information about training data to the risk of leaking confidential data in the cloud. The authors discuss various security techniques, including anonymisation, homomorphic encryption, federated learning, and differential privacy, highlighting their advantages and limitations. The key conclusion is that there is no perfect solution – each method has its drawbacks, and adequate data protection depends on the specific use case. The article also draws attention to regulations such as the GDPR, which enforce privacy, and emphasises that research is still needed to improve existing security technologies.

Muhlhoff et al. in [18] considered the risks of so-called “predictive privacy”. The author points out that modern AI and big data analytics systems can predict sensitive information about individuals (for example, health, financial status, sexual orientation) even if the person has never shared this data. The problem is that current data protection regulations, such as GDPR, do not sufficiently regulate anonymous data and predictive models, allowing large technology companies to use predictions in ways that lead to discrimination and social inequality. The authors propose a new approach to data protection that focuses on collective responsibility and the regulation of the “predictive power” of corporations, rather than just the rights of individuals to protect their data.

Liang et al. in [16] addressed the issue of providing personal data privacy in blockchain-based systems. They discovered that existing solutions suffer from numerous limitations, including a lack of user control over the information, a lack of transparency, and challenges in assigning access rights to the data. Based on this, they presented a novel system, PDPChain, built upon a consortium blockchain that restricts access to designated parties. The authors employed a hybrid approach that leveraged off-chain and on-chain storage solutions; more specifically, the encrypted content is hosted externally to the blockchain, yet access control is enforced on the blockchain. Their approach enables more effective and secure management of personal data on the blockchain, with enhanced user control, while simultaneously optimising system performance.

Abdullahi et al. in [1] drew attention to the threats associated with attacks on biometric systems, which are often overlooked or underestimated. These are mainly attacks on software and biometric templates (i.e., saved patterns of fingerprints, irises, and faces), which can lead to the theft of biometric data. Unlike passwords, these are immutable and irreversible. The authors pointed out that although specific protection mechanisms already exist, universal, effective, and efficient solutions that simultaneously secure biometrics without compromising the quality of the system’s operation are still lacking.

Joyce et al. in [13] examined cross-sectoral data sharing (CDS) in smart cities and the obstacles related to data protection laws (DPL). The study included a literature review and interviews with 14 experts from various smart cities worldwide. The result of this work is the development of a CDS framework model

that facilitates the integration of data sharing with legal requirements. CDS is a key element of innovative city development, but various barriers, including data quality, standardisation, interoperability and privacy issues, complicate its implementation. Data protection laws (DPL) make it difficult to share information, especially when it concerns sensitive data from different sectors (for example, health, transport, and energy). There is a lack of coherent strategies and guidelines that would facilitate the combination of data sharing and privacy protection; therefore, it was necessary to create a CDS framework model. The research results are based on interviews with experts from cities such as Helsinki, Vienna, London, Stockholm, Dubai, Delhi and Bangalore. This model enables the evaluation of the progress levels of innovative city initiatives. The study presented herein offers thoughtful remarks on the direction of future smart cities. The CDS model presented herein facilitates better municipal data governance, allows cross-sector collaboration, and concurrently mandates data protection requirements. This is necessary to develop more integrated, secure, and efficient urban areas of the future.

Pool et al. in [20] examined the problem of health data breaches, which pose a serious challenge to both healthcare providers and patients. The authors conclude that health data breaches are a hidden threat to the digital transformation of the healthcare sector. Without appropriate protection measures, they can undermine patient trust and the stability of healthcare institutions. The most important result of the study is an integrative model that explains the various aspects of health data breaches, including their causes, risk factors and consequences, provides 11 key propositions that help better understand and manage threats, provides a tool for security managers to help them analyze risks and take preventive actions, and emphasizes the need for multi-level analysis of breaches, taking into account technical, organizational and human aspects. The authors concluded that modern patient data protection requires a holistic approach - considering technology, regulations and human factors. Further research should focus on developing innovative protection methods and improving security incident response strategies.

Natarajan et al. in [19] introduced an innovative method of anonymising large medical datasets in the cloud using Scalable Subtree L-Anonymisation (SS-LA) and the Whale Optimisation Algorithm (WOA). The work primarily focuses on privacy protection in electronic health records (EHRs) and evaluates the method's performance based on memory consumption, accuracy, error value, and execution time analysis. SS-LA is an approach based on classical L-Anonymisation, which ensures that each data group contains at least L identical records. This minimises the risk of identifying a specific person. WOA is a meta-heuristic optimisation algorithm inspired by the hunting behaviour of humpback whales (so-called "bubble-net feeding"). It is mainly used in optimisation problems, e.g., parameter selection of machine learning algorithms. Combining SS-LA with WOA offers scalability and efficiency, enabling the fast anonymisation of extensive medical datasets while maintaining usability. WOA optimises the data grouping process, which reduces processing time and minimises anonymisation

errors. There is, however, a lack of privacy protection because such a method does not prevent an effective re-identification attack or guarantee resistance to data pattern analysis. Furthermore, being a heuristic algorithm, WOA does not always yield a globally optimal solution, and the absence of comparison with other state-of-the-art techniques, such as differential privacy, makes it difficult to judge its ability against competing approaches.

Szymoniak et al. in [29] proposed a new method for user verification in IoT environments to generate symmetric keys for securing communication between them. For identity verification, users are identified by facial images analysed by coordinates created by three points: the corners of the eyes and the chin. An alphanumeric session key is generated based on these coordinates and additional time factors. The proposed method has the advantage of eliminating the need to contact the server to generate keys, thereby increasing the system's security and scalability. Studies have shown that the method is effective in recognising users and provides resistance to brute-force attacks. It took over a billion years to break the generated keys, making them exceptionally secure.

Additionally, the article highlights the significance of image resolution and facial angles, which impact the accuracy of the coordinates and the generated key. Solutions to improve the quality of verification are also suggested, such as placing a frame indicating the appropriate face position. The advantages of this system also include eliminating the need for key management by a central server, which reduces the risk of security breaches. It can also be used in various IoT environments, such as medical devices or industrial settings, where computing resources may be limited.

5 Future Challenges and Research Directions

One of the biggest challenges in personal data protection is the dynamic development of technology, which often outpaces applicable legal regulations. Regulations such as GDPR or CCPA were created when artificial intelligence, blockchain, or the Internet of Things were not as prevalent as they are today. There is a need to adapt the law to new realities, including better protection of users from invasive data collection and analysis practices. Future research should focus on privacy protection methods in the context of modern technologies and on developing flexible regulations that will not inhibit innovation, while also ensuring user safety.

With growing cyberattacks, human beings realise that artificial intelligence works to protect data. Machine learning solutions are being used to detect anomalous activity and unauthorised access attempts, but their success relies on the quality and variability of the training data used. The problem is the so-called “black box” of AI - the lack of transparency in the operation of these systems, which can lead to erroneous decisions, such as blocking legal operations or overlooking real threats. Future research should focus on developing explainable AI models that enable a deeper understanding of decision-making mechanisms and enhance trust in automated data protection systems.

A further important study area is data protection in decentralised systems, including blockchain technology. While such technology is more secure and more complicated to manipulate, it raises new challenges, such as how the right to be forgotten can be enforced when information is irretrievably stored in distributed ledgers. Coming up with a way to keep the data safe while also being capable of deleting it when needed will be very important. We must combine technologies like artificial intelligence, cryptography, and blockchain. This will enable us to create end-to-end data protection solutions that encrypt data and comply with relevant laws.

6 Conclusions

At a time when privacy has become the new currency of the digital economy, securing such data has become one of the most significant challenges, involving multiple approaches. The present study critically discussed the serious threats against personal data, the effectiveness of existing protection mechanisms, and potentially future technologies that will shape future strategies for data security.

The outcomes of this research demonstrate that personal data protection is a complex combination of legal and technical issues. Given that the capabilities for compliance with GDPR-like regulations are among the most critical components of protecting personal data and that they have not yet evolved, organisations need to invest substantially in technical measures whose foundation will include encryption, data anonymisation, and AI threat detection. They also need to implement proactive strategies such as intrusion detection systems and regular cybersecurity audits to be prepared for the inevitable evolution of cyber threats.

Emerging technologies, such as AI and blockchain, provide both opportunities and challenges for personal data security. AI enhances security by enabling the real-time detection of potential threats, but it also carries certain risks, such as adversarial attacks. In the future, blockchain technology may be further developed to address issues of personal privacy and scalability, as these currently limit its potential for widespread application. Integration with existing regulatory frameworks will be important in ensuring the safe deployment of these technologies in future research endeavours.

Emphasising user education is another vital part of data protection. Sophistication in security technology will increasingly be checked by human errors, which are significant vulnerabilities. With awareness created through training programs and the implementation of friendly human practices within a highly secure system, it can be effectively reduced in terms of legality.

Looking ahead, the rapid advancement of quantum computing and IoT devices will create a new kind of security challenge. Future digital ecosystems will require quantum-safe encryption techniques and more enhanced and secure models for interconnected smart devices.

An integrated and dynamic approach ensures that the legal, technological, and human components work together to safeguard personal information holistically. In this light, the paper makes a significant contribution to the discussion

by providing a structured framework for considering current and emerging challenges posed by data protection. The evolving protection strategies, in collaboration with policymakers, empower organisations to develop even newer, more secure, and privacy-respecting digital environments.

References

1. Abdullahi, S.M., Sun, S., Wang, B., Wei, N., Wang, H.: Biometric template attacks and recent protection mechanisms: a survey. *Inf. Fusion* **103**, 102144 (2024)
2. Birthriya, S.K., Ahlawat, P., Jain, A.K.: Detection and prevention of spear phishing attacks: a comprehensive survey. *Comput. Secur.* 104317 (2025)
3. Darapu, K., Marukukula, M., et al.: Fraud detection and prevention in finance and banking using artificial intelligence. In: *Real-World Applications of AI Innovation*, pp. 213–232. IGI Global Scientific Publishing (2025)
4. El Mestari, S.Z., Lenzini, G., Demirci, H.: Preserving data privacy in machine learning systems. *Comput. Secur.* **137**, 103605 (2024)
5. Fabrègue, B.F., Bogoni, A.: Privacy and security concerns in the smart city. *Smart Cities* **6**(1), 586–613 (2023)
6. GDPR, G.D.P.R.: General data protection regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (2016)
7. Hirano, M., Kobayashi, R.: Ransmap: open dataset of ransomware storage and memory access patterns for creating deep learning based ransomware detectors. *Comput. Secur.* **150**, 104202 (2025)
8. Hirschey, J.K.: Symbiotic relationships: pragmatic acceptance of data scraping. *Berkeley Tech. LJ* **29**, 897 (2014)
9. Hossain, S.T., Yigitcanlar, T., Nguyen, K., Xu, Y.: Cybersecurity in local governments: a systematic review and framework of key challenges. *Urban Governance* (2025)
10. Humayun, M., Tariq, N., Alfayad, M., Zakwan, M., Alwakid, G., Assiri, M.: Securing the internet of things in artificial intelligence era: a comprehensive survey. *IEEE Access* (2024)
11. Huseynov, E.: Identity theft. In: *Computer and Information Security Handbook*, pp. 993–1006. Elsevier (2025)
12. Javaid, M., Haleem, A., Singh, R.P., Suman, R.: Towards insighting cybersecurity for healthcare domains: a comprehensive review of recent practices and trends. *Cyber Secur. Appl.* **1**, 100016 (2023)
13. Joyce, A., Javidroozi, V.: Smart city development: data sharing vs. data protection legislations. *Cities* **148**, 104859 (2024)
14. Khanna, A.: Ransomware lifecycle. In: *Securing an Enterprise: Maximizing Digital Experiences through Enhanced Security Measures*, pp. 105–118. Springer, Cham (2025)
15. Kumar, S., Dwivedi, M., Kumar, M., Gill, S.S.: A comprehensive review of vulnerabilities and AI-enabled defense against DDOS attacks for securing cloud services. *Comput. Sci. Rev.* **53**, 100661 (2024)
16. Liang, W., et al.: Pdpchain: a consortium blockchain-based privacy protection scheme for personal data. *IEEE Trans. Reliab.* **72**(2), 586–598 (2022)

17. McIntosh, T., et al.: Ransomware reloaded: re-examining its trend, research and mitigation in the era of data exfiltration. *ACM Comput. Surv.* **57**(1), 1–40 (2024)
18. Mühlhoff, R.: Predictive privacy: collective data protection in the context of artificial intelligence and big data. *Big Data Soc.* **10**(1), 20539517231166890 (2023)
19. Natarajan, A., Shanthi, N.: Optimizing healthcare big data privacy with scalable subtree-based l-anonymization in cloud environments. *Wirel. Netw.* 1–16 (2025)
20. Pool, J., Akhlaghpour, S., Fatehi, F., Burton-Jones, A.: A systematic analysis of failures in protecting personal health data: a scoping review. *Int. J. Inf. Manage.* **74**, 102719 (2024)
21. Prümmer, J., van Steen, T., van den Berg, B.: Assessing the effect of cybersecurity training on end-users: a meta-analysis. *Comput. Secur.* **150**, 104206 (2025)
22. Roy, P.K., Kumar, P., Bhattacharya, A.: ZeroVCS: an efficient authentication protocol without trusted authority for zero-trust vehicular communication systems. *Futur. Gener. Comput. Syst.* **163**, 107520 (2025)
23. Saeed, S., Altamimi, S.A., Alkayyal, N.A., Alshehri, E., Alabbad, D.A.: Digital transformation and cybersecurity challenges for businesses resilience: issues and recommendations. *Sensors* **23**(15), 6666 (2023)
24. Schmitz-Berndt, S.: Defining the reporting threshold for a cybersecurity incident under the NIS directive and the NIS 2 directive. *J. Cybersecur.* **9**(1), tyad009 (2023)
25. Singh, N., Do, Y., Yu, Y., Fouad, I., Kim, J., Kim, H.: Crumbled cookies: exploring e-commerce websites' cookie policies with data protection regulations. *ACM Trans. Web* **19**(1), 1–24 (2025)
26. Szymoniak, S.: Using a security protocol to protect against false links. In: *Moving Technology Ethics at the Forefront of Society, Organisations and Governments*, pp. 513–525. Universidad de La Rioja (2021)
27. Szymoniak, S.: Key distribution and authentication protocols in wireless sensor networks: a survey. *ACM Comput. Surv.* **56**(6), 1–31 (2024)
28. Szymoniak, S., Depta, F., Karbowski, Ł., Kubanek, M.: Trustworthy artificial intelligence methods for users' physical and environmental security: a comprehensive review. *Appl. Sci.* **13**(21), 12068 (2023)
29. Szymoniak, S., Kubanek, M.: Biometry-based verification system with symmetric key generation method for internet of things environments. *Sci. Rep.* **15**(1), 5464 (2025)
30. Szymoniak, S., Siedlecka-Lamch, O., Kurkowski, M.: Sat-based verification of NSPK protocol including delays in the network. In: *2017 IEEE 14th International Scientific Conference on Informatics*, pp. 388–393. IEEE (2017)
31. Varshney, G., Kumawat, R., Varadharajan, V., Tupakula, U., Gupta, C.: Anti-phishing: a comprehensive perspective. *Expert Syst. Appl.* **238**, 122199 (2024)



Deepfake Manipulation and Ethical Dilemmas: A Comprehensive Risk Assessment

Mariusz Kubanek and Sabina Szymoniak

Department of Computer Science, Czestochowa University of Technology,
Dabrowskiego 69, Czestochowa, Poland
{mariusz.kubanek,sabina.szymoniak}@icis.pcz.pl

Abstract. Deepfake technology, based on artificial intelligence (AI), enables the creation of realistic images, videos, and text, bringing new possibilities and posing serious ethical, social, and legal threats. The article examines key areas in which deepfakes affect contemporary society. This technology can be used to create false materials aimed at social destabilisation and propaganda, including disinformation. In the area of cyberbullying, deepfakes contribute to privacy violations such as pornography and blackmail. Additionally, generating false evidence can affect court decisions and facilitate financial fraud. This problem also affects the crisis of public trust, undermining the media's credibility. The authors emphasise the need to introduce safeguards such as watermarking, public education and identification of AI-generated content. The need for international legal regulations and the development of technologies to detect deepfakes is indicated. The article considers ethical issues associated with this technology, minimising the risk of abuse and supporting innovation sustainably.

Keywords: deepfake · ethics · disinformation · image and text manipulation

1 Introduction

The development of deepfake technology has become one of the most controversial achievements of artificial intelligence in recent years. The ability to generate realistic images, videos, and texts has opened up new perspectives for entertainment, education, and business. Thanks to tools that utilise neural networks and machine learning algorithms, it has become possible to create materials that, at first glance, appear authentic but are entirely synthetic [37, 42].

However, alongside innovative applications, serious threats have emerged. Deepfakes are increasingly used to manipulate information, spread disinformation, cyberbullying or financial fraud. Fake political recordings can influence public opinion, manipulated video materials can become a tool for blackmail, and the generated voices allow cybercriminals to impersonate real people. In a

world where digital content plays a key role in communication, deepfakes undermine the foundations of trust in the media and institutions [14,47]. Also, many other methods for collecting and performing data support deepfake technology [33,41].



Fig. 1. Summary of domains where deepfake technology can be used [1].

Figure 1 identifies active and potential hazards associated with deepfake technology. It is an excellent tool for manipulating visual and audio data to create highly realistic fake material for various factors. The primary concern about how deepfakes could influence people's opinions on non-substantive political speeches by political leaders or even interviews is through electoral manipulation. Social engineering is another threat that modified content can be used to fake scams and cyberattacks, making them appear real. Most likely, automated disinformation campaigns are another menace, distributing falsehoods online. Deepfakes are also used for identity theft purposes by falsifying documents or impersonating other individuals. In finance, they can be used to blackmail individuals using mimicked video calls and audio. Deepfake-driven scams and hoaxes undermine the trustworthiness of online content, and pornographic videos of public personalities are an invasion of privacy and a critical risk for victims. The magnitude of such risks underscores the need to develop effective mechanisms for detecting and preventing deepfake abuses [2,6,13,23].

This article aims to analyse the risks associated with the development of this technology and to identify ethical and technical solutions for its safe use.

Key risk areas and methods for mitigating them will be discussed through legal regulations and technical tools that enable the detection and prevention of abuse. In the face of increasing digitalisation and increasingly advanced methods of manipulation, it is necessary to develop effective strategies to balance innovation and social responsibility.

2 Deepfake Technology: How It Works

The study of artificial intelligence that incorporates algorithms that learn from data without being explicitly programmed is known as machine learning. This enables a computer to analyse vast amounts of data, find patterns and make decisions without being explicitly programmed. Depending on the approach, there are many ways to classify machine learning. Supervised learning is a technique where an algorithm is trained on input data along with the correct output, allowing it to predict the output of new samples. Unsupervised learning occurs when the model works on the data without knowing the labels and can discover structure without being explicitly instructed on what to look for. Yet another approach is Reinforcement Learning, where an agent interacts with an environment to learn how to act optimally in a given condition by receiving a reward for a good action and a penalty for a wrong action [7, 43, 52].

The sub-discipline of the largest branch of Machine Learning, Deep Learning, is inspired by the human brain's neural networks. Neurons are many connected layers of increasing complexity that learn to transform raw data into more meaningful representations. They are the first layers that learn features like shapes and colours, and then other layers learn to recognise patterns, leading to the identification of objects, faces, or even sounds. Therefore, it is possible to ascribe the development of the theory of deep learning to the development of the machine learning theory to solve such complex problems as image recognition, language transformation, and even the creation of genuine content [9, 34].

Generative adversarial networks, known as GANs, are one of the most significant achievements of deep learning. These are models consisting of two neural networks that learn through competition. The first one, the generator, creates artificial images, recordings or sounds that are supposed to resemble real ones. The second one, the discriminator, assesses their authenticity and attempts to distinguish between fake and real content. During training, both networks improve each other: the generator creates increasingly realistic materials, and the discriminator becomes better at detecting manipulations. Thanks to this, GAN enables the generation of images and videos that are almost indistinguishable from the real ones [15, 30].

Deepfake technology utilises GAN models to generate convincing yet artificially created video and audio content. Advanced algorithms enable accurate reproduction of facial expressions, voices, and mouth movements, making fake recordings look incredibly realistic. This type of technology has found applications in various areas, including entertainment, education, cybersecurity [50, 51], and information manipulation. It enables digital rejuvenation of actors, reconstruction of historical figures, and real-time speech translation, but at the same

time, it can be used for disinformation, fraud, and privacy violations. Developing deepfakes requires effective detection methods and robust legal regulations to mitigate the risk of abuse and enhance security in the digital world [15,30]. Figure 2 summarises the Generative Adversarial Network operation.

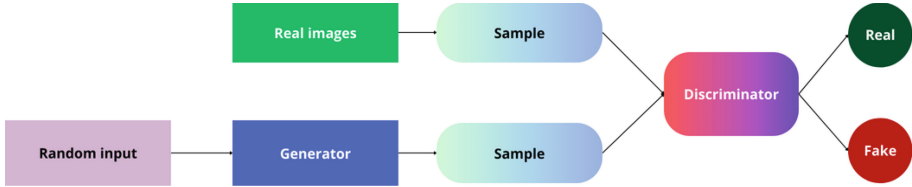


Fig. 2. Generative Adversarial Network operation [Source: self-elaboration].

Deepfakes are created with the help of advanced artificial intelligence technologies, such as GANs, Machine Learning, and Deep Learning. Refining the generated content using GANs makes it harder to distinguish between real and fake content. However, machine learning models, such as autoencoders, are also used to create deepfakes, which analyse and reconstruct images to change the face and voice in videos. The deep learning models provide a realistic replication of facial and mouth movements in sync with the generated sound of the text provided as input. Table 1 shows examples of positive (P) and negative (N) deepfakes applications.

As mentioned earlier, GANs are one of the techniques used for deepfake detection. Tchaptchet et al. in [45] focused on detecting fake content using eye irises. They applied a gradient map to the iris; the method visualised biological characteristics, such as round shape, identical reflections, and iris size.

On the other hand, Talib et al. in [44] noticed that deepfake is a rapidly evolving attack technique that involves synthesising thousands of photorealistic images or manipulating large videos. Additionally, they observed that existing methods, such as GAN, are vulnerable to the spread of deepfakes. They proposed two novel solutions, ColorDense and LightDense, which utilise the DenseNet network backbone to distinguish between fake and real face images. These models incorporated luminance and colour signals from multi-colour spaces, extracting important features and passing them to a classification system using neuronal techniques.

The positive and negative uses of deepfakes demonstrate the power and danger this technology poses. Legal regulations and effective detection methods are required to limit the risk of abuse [27,49]. Smart devices enhance multimedia content creation, but with the advancement of deep learning techniques, creating sophisticated deepfake content, including speech deepfakes, has become easier. To counter this threat, researchers have proposed speech Deepfake detection methods [22].

Table 1. Positive and negative deepfakes’ applications.

Area	Description	P/N
Film and entertainment industry	Deepfakes allow for the digital rejuvenation of actors, their post-mortem reenactment, and reduced special effects costs	P
Education, training	Deepfakes can create realistic simulations of historical figures that “speak” to audiences, enriching teaching	P
Medicine and therapy	Deepfakes can help people who have lost their voice by generating synthetic speech based on their previous recordings	P
Art and creativity	Artists use deepfakes to create new forms of expression, combining elements of different visual and audio styles	P
Security and privacy	Deepfakes can be used to protect identities, such as generating anonymous faces for informants	P
Political manipulation	Fake recordings can be used to discredit politicians, influence elections, and spread disinformation	N
Financial fraud	Cybercriminals use deepfakes to impersonate company executives to extort funds (known as Business Email Compromise)	N
Fake pornography	Creating manipulated content featuring public and private figures without their consent poses a serious privacy risk	N
Scams, online fraud	Deepfakes can make fake phone calls or recordings look authentic, making phishing attacks more effective	N
Undermining trust in the media	In an era of disinformation, deepfakes can lead to authentic content losing credibility and spreading false narratives	N

3 Key Risks of Deepfake Technology

Deepfakes are a technology that allows for the creation of incredibly realistic fake recordings and photos. Although they can have positive applications, such as in the film industry or education, they are unfortunately often used unethically. Here are the most significant threats associated with them:

Deepfakes have become powerful tools for manipulation. Fake recordings of politicians can influence election results, and doctored videos compromising famous people can tarnish their reputations in an instant. In the age of social media, this type of content spreads rapidly, and the average recipient often lacks the tools to distinguish truth from manipulation [8, 40].

One of the most common and harmful uses of deepfakes is using them to create fake pornographic materials. Women are often the victims, whose faces are edited into such videos without their knowledge or consent. It is not only a violation of privacy, but often a form of blackmail and psychological harassment [12].

For example, in Santoro et al. [32], an AI tool is proposed that anyone can use to identify bullying behaviour and prevent cyberbullying. The tool utilises surveys to categorise and tag individuals, providing teachers with valuable insights into online bullying incidents. The framework includes modules for detecting physical violence, deepfake recognition, and behavioural modelling. The tool prioritises transparency and explainability for informed intervention.

Deepfakes are not only an image problem but also a real threat to companies and institutions. Cybercriminals can impersonate CEOs of large companies, issuing false transfer orders. There is also a risk that manipulated recordings will be submitted to courts as “evidence”, which can lead to unfair sentences or make it difficult to prosecute real criminals [31].

Khan et al. in [18] examined the impact of deepfake technology on the financial sector, highlighting its potential to disrupt authenticity and erode confidence. It examines detection efforts, transaction weaknesses, market manipulation, and corporate communications. The research combines industry, regulatory, and academic perspectives to provide practical techniques for reducing risks and preserving financial system integrity.

It used to be said that “if something is on the recording, it means it is true”. Today, this is no longer true. It is increasingly difficult to believe in the authenticity of any video material. This leads to a dangerous phenomenon - even real recordings can be questioned as deepfakes, which weakens trust in the media, justice and public institutions [16,35].

Deepfake is a technology that is getting out of control. As it becomes more advanced, detection becomes increasingly complex, and the consequences of its abuse can be catastrophic. That is why it is essential to develop tools to identify fake recordings and educate society about the dangers of this technology.

4 Ethical and Social Perspective

The rise of deepfake technology raises more questions about who is responsible and what happens when people use it. The ability to create fake audio and video that appear real opens new doors for entertainment, education, and business. But it also means there’s a chance for people to trick others, scam them, or invade their privacy [10,21,29].

Should the creators of deepfake algorithms be responsible for how their technologies are used? Perhaps too much blame is placed on users who exploit hijack situations by knowingly creating false information that could mislead and blackmail others as well. This is challenging because some of these tools can be useful for critical purposes, such as laws that do good, but they can also be misused for malicious purposes. The problem is not the technology but the lack of mechanisms to control its abuse [17].

Deepfakes can be used in film productions, computer games or historical reconstructions. Thanks to them, it is possible, for example, to “bring” actors back to life on screen or create realistic dubbing in different languages. However, where manipulation begins, a problem arises. An ethical boundary is crossed if a

deepfake is used to create a false statement by a politician or to deceive someone in a video call. There are still no clear regulations defining what is acceptable and what is not [19].

As deepfakes become increasingly challenging to detect, educating the public is becoming crucial. We should learn to approach online content critically and use methods to verify information. It is also worth developing technologies that will help identify manipulated materials. Only a combination of public awareness and tools that detect forgeries can reduce the scale of threats associated with deepfakes [28].

Soto et al. in [39] demonstrated the effectiveness of a narrative processing model in explaining the psychological relationship between receivers and various types of deepfakes, regardless of their genre. However, the results contradict previous suggestions that the narrative genre provides cues for evaluating plausibility. The study also found that the more receivers enjoy a deepfake, the greater their intention to share it on social networks. Additionally, they found a relationship between specific aspects of the psychological processing of deepfakes (familiarity with the person depicted, narrative transportation, and enjoyment) and the intention to share them on social media networks. This highlights the importance of understanding the psychological mechanisms that explain the effect of narrative processing on sharing behaviour.

Shao et al. in [36] found out that deepfake detection methods often struggle to generalise to unseen or degraded samples due to over-fitting of low-level forgery patterns. They proposed the first parameter-efficient tuning approach, DeepFake-Adapter, which efficiently adapts high-level semantics from large pre-trained Vision Transformers (ViTs) [48] for deepfake detection. This approach introduced lightweight dual-level adapter modules to a ViT, guiding the adaptation process to be aware of global and local forgery cues.

Deepfakes are a technology that is neither inherently good nor bad in itself. It is up to people to decide how it will be used—whether it will become a tool for fraud and manipulation or support for creativity and innovation.

5 Protection and Countermeasures

Deepfakes are becoming an increasingly advanced and accessible technology, which raises the need to develop effective methods of defending against their misuse. In response to the threats, technological tools, legal solutions, and strategies for international cooperation are being developed to address these challenges. Deepfake Detection Technologies.

Scientists and technology companies are developing increasingly sophisticated detection algorithms to identify counterfeit materials. They use AI, among other things, to analyse subtle anomalies in facial expressions, eye movements, or image shading. Popular techniques include anomaly analysis, searching for unnatural patterns in recordings, and watermarking or marking content in a way that allows for verification of its authenticity. However, these tools must constantly evolve because deepfakes are becoming increasingly realistic.

The first regulations regarding deepfakes have appeared in many countries, although their effectiveness remains controversial. In the United States, some states have already introduced bans on creating deepfakes in the context of elections, and the European Union is working on regulations regarding liability for generating false materials. However, the boundary between creative freedom and public safety remains a problem— not every modification of an image or voice has to be treated as a threat [24,38].

Fighting deepfakes requires the involvement of governments, technology companies and international organisations. Google, Microsoft, and Meta continue to research new applications of the art of detecting potential manipulations, while Europol and Interpol investigate cases of abuse related to deepfakes in cyber-crime. Public awareness is also critical: the more people know how to identify fake content, the harder it will be for fraudsters to wield it for manipulation [3].

6 Future of Deepfake: Opportunities and Threats

Deepfake is a technology that, on the one hand, opens up new possibilities in art, entertainment and education, and on the other hand, raises serious concerns about disinformation and fraud. As it develops, new methods of manipulation and more advanced methods of detecting fake content will appear. Development of technology and new methods of manipulation [11,25].

Deepfakes are becoming increasingly difficult to detect. Modern algorithms can generate realistic video and audio that do not reveal typical errors, such as unnatural blinking or image artefacts. Models based on generative neural networks (GANs) are constantly improving. In the future, deepfakes may reach a level where distinguishing a fake recording from a real one will become almost impossible with the naked eye [26,46].

The metaverse, which is comprised of virtual worlds based on VR and AR, may become a new space for deepfakes. Since it is already possible to generate realistic avatars today, deepfakes may enable the impersonation of other people in real-time, which poses a serious threat to users' privacy and security. How can we protect people from fraud when digital identity can be easily manipulated?

Not everything associated with deepfakes has to be negative. This technology can be used positively, for example, in films and games to recreate the image of actors or in education to simulate historical events. Medical applications are also possible in communication for people who have lost the ability to speak. However, it will be crucial to create clear ethical rules and regulations that allow deepfakes to be used safely and responsibly [5,20]. Alanazi et al. in [4] emphasised the importance of ethically guided approaches and interdisciplinary collaboration in policy responses. It provides evidence-based recommendations for mitigating risks and harnessing potential benefits.

Therefore, the future of deepfakes depends on how humanity approaches this technology—whether it will be a tool for creative expression or a weapon for disinformation and deception.

7 Conclusions

Deepfake technology is an innovation that simultaneously opens up new possibilities and poses serious threats. As this article demonstrates, the use of deepfakes spans a range of applications, from creative uses in art and education to more concerning instances of information manipulation, fraud, and privacy violations.

The impact of deepfakes on society is particularly worrying. In an era of rapid digitisation and the vast amount of content published online, fake recordings can effectively undermine trust in the media, institutions, and even personal relationships. Political disinformation, financial fraud, and cyberbullying are just some of the consequences we are already seeing.

On the contrary, it is essential to remember that, like any tool, deepfakes are a technology that users can use for good or evil. Some of the possible legitimate uses of this technology may include reconstructing historical artefacts, films, or educational materials.

It is necessary to address the issue on multiple levels to mitigate its adverse effects. Developing AI-based deepfake detection systems, implementing digital content marking (such as watermarking), and monitoring and recording anomalies are essential technological steps. However, legal regulations are equally important, and they should clearly define the limits of liability and penalties for abuse. Learning critical thinking and becoming socially aware of manipulated content will help reduce the impact of deepfakes on everyday life. In the future, we anticipate further advancements in deepfake technology, including the generation and detection of realistic content. Will society manage to keep up with the progress made in deepfake technology? To ensure the gains achieved from the advancement of this technology and reduce its risks, the responsibility of regulation, investment in protective tools, and building user awareness are essential foundations.

References

1. What is deepfake: AI endangering your cybersecurity? (2025). <https://www.fortinet.com/resources/cyberglossary/deepfake>. Accessed 07 Jan 2025
2. Akhtar, Z., Pendyala, T.L., Athmakuri, V.S.: Video and audio deepfake datasets and open issues in deepfake technology: being ahead of the curve. *Forensic Sci.* **4**(3), 289–377 (2024)
3. Alanazi, S., Asif, S.: Exploring deepfake technology: creation, consequences and countermeasures. *Hum.-Intell. Syst. Integr.* 1–12 (2024)
4. Alanazi, S., Asif, S., Caird-daley, A., Moulitsas, I.: Unmasking deepfakes: a multi-disciplinary examination of social impacts and regulatory responses. *Hum.-Intell. Syst. Integr.* 1–23 (2025)
5. Bendiab, G., Haiouni, H., Moulas, I., Shiales, S.: Deepfakes in digital media forensics: generation, AI-based detection and challenges. *J. Inf. Secur. Appl.* **88**, 103935 (2025)
6. Cai, Z., et al.: AV-Deepfake1M: a large-scale LLM-driven audio-visual deepfake dataset. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7414–7423 (2024)

7. Chander, B., John, C., Warriar, L., Gopalakrishnan, K.: Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness. *ACM Comput. Surv.* **57**(6), 1–49 (2025)
8. Chapagain, D., Kshetri, N., Aryal, B.: Deepfake disasters: a comprehensive review of technology, ethical concerns, countermeasures, and societal implications. In: 2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), pp. 1–9. IEEE (2024)
9. Chen, X., et al.: Deep learning-based software engineering: progress, challenges, and opportunities. *SCIENCE CHINA Inf. Sci.* **68**(1), 1–88 (2025)
10. Edwards, P., Nebel, J.C., Greenhill, D., Liang, X.: A review of deepfake techniques: architecture, detection and datasets. *IEEE Access* (2024)
11. Gambín, Á.F., Yazidi, A., Vasilakos, A., Haugerud, H., Djenouri, Y.: Deepfakes: current and future trends. *Artif. Intell. Rev.* **57**(3), 64 (2024)
12. Gupta, R., Singh, A.K., Mittal, P., et al.: AI based cyberbullying detection and prevention. In: 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON), pp. 1–6. IEEE (2024)
13. Heidari, A., Jafari Navimipour, N., Dag, H., Unal, M.: Deepfake detection using deep learning methods: a systematic and comprehensive review. *Wiley Interdisc. Rev. Data Mining Knowl. Discov.* **14**(2), e1520 (2024)
14. Javed, M., Zhang, Z., Dahri, F.H., Laghari, A.A.: Real-time deepfake video detection using eye movement analysis with a hybrid deep learning approach. *Electronics* **13**(15), 2947 (2024)
15. Kanniga Devi, R., Asha Jerlin, M.: Generative adversarial networks: a comprehensive review. In: *Generative Artificial Intelligence (AI) Approaches for Industrial Applications*, pp. 69–92 (2025)
16. Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., Xia, F.: Deepfake video detection: challenges and opportunities. *Artif. Intell. Rev.* **57**(6), 159 (2024)
17. Kaur, A., Noori Hoshyar, A., Wang, X., Xia, F.: Beyond deception: exploiting deepfake technology for ethical innovation in healthcare. In: *Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine*, pp. 70–78 (2024)
18. Khan, R., Taqi, M., Afzal, A.: Deepfakes in finance: unraveling the threat landscape and detection challenges. In: *Navigating the World of Deepfake Technology*, pp. 91–120. IGI Global (2024)
19. Kubanek, M., Szymoniak, S.: Ethical challenges in AI integration: a comprehensive review of bias, privacy, and accountability issues. *The Leading Role of Smart Ethics in the Digital World*, pp. 75–85 (2024)
20. Kumar, A., Singh, D., Jain, R., Jain, D.K., Gan, C., Zhao, X.: Advances in deepfake detection algorithms: exploring fusion techniques in single and multi-modal approach. *Inf. Fusion* 102993 (2025)
21. Lee, G., et al.: Dual-channel deepfake audio detection: leveraging direct and reverberant waveforms. *IEEE Access* (2025)
22. Li, M., Ahmadiadli, Y., Zhang, X.P.: A survey on speech deepfake detection. *ACM Comput. Surv.* (2025)
23. Maheshwari, R.U., Paulchamy, B., Pandey, B.K., Pandey, D.: Enhancing sensing and imaging capabilities through surface plasmon resonance for deepfake image detection. *Plasmonics* 1–20 (2024)
24. Mania, K.: Legal protection of revenge and deepfake porn victims in the European union: findings from a comparative legal study. *Trauma Violence Abuse* **25**(1), 117–129 (2024)

25. Maniyal, V., Kumar, V.: Unveiling the deepfake dilemma: framework, classification, and future trajectories. *IT Prof.* **26**(2), 32–38 (2024)
26. Montasari, R.: Responding to deepfake challenges in the united kingdom: Legal and technical insights with recommendations. In: Montasari, R. (ed.) *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses*, pp. 241–258. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-50454-9_12
27. Nadimpalli, A.V., Rattani, A.: Proactive deepfake detection using GAN-based visible watermarking. *ACM Trans. Multimed. Comput. Commun. Appl.* **20**(11), 1–27 (2024)
28. Navarro Martínez, O., Fernández-García, D., Cuartero Monteagudo, N., Forero-Rincón, O.: Possible health benefits and risks of deepfake videos: a qualitative study in nursing students. *Nurs. Rep.* **14**(4), 2746–2757 (2024)
29. Premkumar, S., Arthi, T., Arora, K., Vinotha, D., Renu, Y.: Deepfake video detection using mouth movement technology. In: Rama Sree, S., Kumar, S. (eds.) *ICACTEA 2024*, pp. 139–144. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72747-4_21
30. Purwono, P., Wulandari, A.N.E., Ma'arif, A., Salah, W.A.: Understanding generative adversarial networks (GANs): a review. *Control Syst. Optim. Lett.* **3**(1), 36–45 (2025)
31. Sandoval, M.P., de Almeida Vau, M., Solaas, J., Rodrigues, L.: Threat of deepfakes to the criminal justice system: a systematic review. *Crime Sci.* **13**(1), 41 (2024)
32. Santoro, E., Marrone, S., Sansone, C.: Explainable AI for bullying and cyberbullying detection. In: *Proceedings of the 2024 Conference on Human Centred Artificial Intelligence-Education and Practice*, p. 59 (2024)
33. Schreiner, W., Steingartner, W., Novitzká, V.: A novel categorical approach to semantics of relational first-order logic. *Symmetry* **12**(10), 1584 (2020). <https://doi.org/10.3390/sym12101584>
34. Shah, M., Sureja, N.: A comprehensive review of bias in deep learning models: methods, impacts, and future directions. *Arch. Comput. Methods Eng.* **32**(1), 255–267 (2025)
35. Shao, R., Wu, T., Liu, Z.: Robust sequential deepfake detection. *Int. J. Comput. Vis.* 1–18 (2025)
36. Shao, R., Wu, T., Nie, L., Liu, Z.: Deepfake-adapter: dual-level adapter for deepfake detection. *Int. J. Comput. Vis.* 1–16 (2025)
37. Shi, L., Zhang, J., Ji, Z., Bai, J., Shan, S.: Real face foundation representation learning for generalized deepfake detection. *Pattern Recognit.* 111299 (2024)
38. Van der Sloot, B., Wagenveld, Y.: Deepfakes: regulatory challenges for the synthetic society. *Comput. Law Secur. Rev.* **46**, 105716 (2022)
39. Soto-Sanfiel, M.T., Angulo-Brunet, A., Saha, S.: Deepfakes as narratives: psychological processes explaining their reception. *Comput. Hum. Behav.* **165**, 108518 (2025)
40. Stanciu, A., Ciupercă, E.M.: Can deepfakes benefit the metaverse in an era of disinformation? Insights from a systematic review. *IFAC-PapersOnLine* **58**(3), 61–65 (2024)
41. Steingartner, W., Novitzká, V.: Categorical model of structural operational semantics for imperative language. *J. Inf. Organ. Sci.* **40**(2), 203–219 (2016)
42. Szymoniak, S., Depta, F., Karbowski, Ł., Kubanek, M.: Trustworthy artificial intelligence methods for users' physical and environmental security: a comprehensive review. *Appl. Sci.* **13**(21), 12068 (2023)

43. Szymoniak, S., Kubanek, M.: Biometry-based verification system with symmetric key generation method for internet of things environments. *Sci. Rep.* **15**(1), 5464 (2025)
44. Talib, M.A., Nasir, Q., Nassif, A.B., Fadhl, N.B., Gouda, O.M.: Chrominance and luminance: a study to detect deepfakes. *Multimedia Tools Appl.* 1–19 (2025)
45. Tchaptchet, E., Tagne, E.F., Acosta, J., Danda, R., Kamhoua, C.: Deepfakes detection by iris analysis. *IEEE Access* (2025)
46. Twomey, J., Ching, D., Aylett, M.P., Quayle, M., Linehan, C., Murphy, G.: What is so deep about deepfakes? A multi-disciplinary thematic analysis of academic narratives about deepfake technology. *IEEE Trans. Technol. Soc.* (2024)
47. Wang, T., Liao, X., Chow, K.P., Lin, X., Wang, Y.: Deepfake detection: a comprehensive survey from the reliability perspective. *ACM Comput. Surv.* **57**(3), 1–35 (2024)
48. Wang, Y., Deng, Y., Zheng, Y., Chattopadhyay, P., Wang, L.: Vision transformers for image classification: a comparative survey. *Technologies* **13**(1), 32 (2025)
49. Wang, Y., Huang, H.: Audio-visual deepfake detection using articulatory representation learning. *Comput. Vis. Image Underst.* **248**, 104133 (2024)
50. Zbrzezny, A.M., Szymoniak, S., Kurkowski, M.: Efficient verification of security protocols time properties using SMT solvers. In: Martínez Álvarez, F., Troncoso Lora, A., Sáez Muñoz, J.A., Quintián, H., Corchado, E. (eds.) *CISIS/ICEUTE - 2019. AISC*, vol. 951, pp. 25–35. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-20005-3_3
51. Zbrzezny, A.M., Zbrzezny, A., Szymoniak, S., Siedlecka-Lamch, O., Kurkowski, M.: Versectis-an agent based model checker for security protocols. In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2123–2125 (2020)
52. Zha, D., et al.: Data-centric artificial intelligence: a survey. *ACM Comput. Surv.* **57**(5), 1–42 (2025)



Digital Agriculture Under Threat: Cybersecurity Challenges and Policy Gaps

Olga Siedlecka-Lamch  

Czestochowa University of Technology, Czestochowa, Poland
olga.siedlecka@icis.pcz.pl

Abstract. Agriculture has undergone a digital revolution in recent years. Advanced technologies such as the Internet of Things (IoT), artificial intelligence, autonomous machines, and cloud data collection have been introduced. We refer to this as Agriculture 4.0 and 5.0. Although such innovations bring measurable benefits in terms of efficiency, adaptability to capabilities, and sustainability, they simultaneously expose agricultural infrastructure to an increasing number of cyber threats. The greater the risks, the less aware the farmers are in this regard. This article provides an overview of research on the connections between cybersecurity and food security, focusing on the analysis of cyberattacks targeting digital agricultural systems. A typology of the main threats, their potential impacts on food supply chains, and a review of current legal regulations have been presented. Based on case studies and international policy, critical security gaps, regulatory and ethical deficiencies have been identified, and directions for further research and practical actions have been proposed. The article emphasizes the necessity of integrated strategies that combine technical, legal, and educational aspects to enhance the resilience of food systems against cyber threats.

Keywords: Cybersecurity in Agriculture · Food Security Risks · Data Security

1 Introduction

The geopolitical situation in the world is changing very dynamically. Equally dynamically, weather conditions are changing. All of this significantly affects the food security of entire communities. Not only those dependent on external aid, but also those groups that have so far lived in prosperity. Food security is the foundation of social peace and prosperity, encompassing both access to food and its quality. According to the definition adopted by the Food and Agriculture Organization of the United Nations (FAO), food security is a state in which: *All people, at all times, have physical, social, and economic access to sufficient, safe, and nutritious food that meets their dietary needs and food preferences for an active and healthy life.* FAO, 1996, World Food Summit [4].

To the factors influencing dynamic changes, another one should be added, often overlooked by many - cybersecurity of modern solutions being introduced

in agriculture. Modern agriculture is undergoing deepening digitization. We have the concepts of Agriculture 4.0 and 5.0 where advanced digital technologies are combined with traditional agronomic knowledge. The Internet of Things (IoT), artificial intelligence, autonomous machines, and cloud data collection already support agricultural production processes today, enabling precise sowing, fertilization, resource management, crop forecasting, and the automation of farm activities. Agriculture 5.0 additionally introduces autonomous systems and emphasizes the importance of sustainable development, system autonomy, and responses to environmental challenges.

With the digital transformation of agriculture, new challenges are emerging. Food security increasingly depends on the stability and resilience of digital systems supporting production and food supply chains. Cyberattacks on agricultural infrastructure can lead to disruptions in fundamental processes such as sowing, irrigation, harvesting, or transportation. Loss of data integrity, communication outages, or attacks on management systems pose a real threat not only to individual farms but to entire social groups.

The largest cyberattack in the agriculture sector to date may be considered the ransomware attack on NEW Cooperative Inc. in the US in 2021. The attack paralyzed the systems managing the grain and feed supply chain and threatened grain and meat supplies. The attack was attributed to the Russian-linked group BlackMatter. The ransomware paralyzed the management of grain handling, feed distribution and logistics. Among the infected software were the company's systems that coordinated livestock feeding schedules, grain storage levels and real-time supply chain planning. The BlackMatter attackers demanded \$5.9 million in ransom and threatened to release stolen data. The cooperative was forced to shut down its systems, disrupting its operations and putting the country's food supply chains, particularly poultry and pork, at risk. The situation exposed the vulnerability of agricultural infrastructure to targeted cyberattacks and how quickly such problems can translate into real threats to food availability. It also exposed the inadequate preparedness of large, technologically advanced farms to defend against complex ransomware attacks [23]. Another example of the real consequences of cyberattacks on modern agriculture comes from Switzerland. At the end of 2023, Vital Bircher, a dairy farmer from Hagendorn (Zug canton), fell victim to a ransomware attack that targeted the software managing his robotic milking system. While the robot was running, access to key data—such as the health status of individual cows, reproductive cycles, and insemination records—was completely blocked. The attackers demanded a ransom of around \$10,000. As a direct result of the lack of access to real-time monitoring and intervention data, one pregnant cow developed complications that went unnoticed; the calf died in utero and the cow had to be euthanized. The farmer reported financial losses of around CHF 6,000–7,000, including veterinary costs and damage to the digital infrastructure. Attacks in the agricultural sector can cause not only economic damage but also real harm [30].

This article aims to highlight the importance of the issue of cyber threats to food security: starting with a review of current research on cyber threats in the

context of digital agriculture and an analysis of their impact on food security, through the typology of attacks and the identification of gaps, to considerations regarding the lack of regulations and legal safeguards. Real-life incident examples were discussed, and directions for further research and potential defense strategies were proposed. The article emphasizes the necessity of a holistic approach to food system security, which will integrate technological, legal, organizational, and ethical aspects.

1.1 Analysis of the Literature on Cyber Threats in Agriculture

The literature review was conducted using a structured approach resembling the principles of a systematic literature review (SLR) together with Elicit tool. Databases used for the analysis included ScienceDirect, Web of Science, and Google Scholar. The search string included keywords such as: (“cybersecurity” OR “cyber threats”) AND (“agriculture” OR “food supply chain”) AND (“IoT” OR “cloud” OR “artificial intelligence”). Boolean operators (AND, OR) were used to combine terms. The initial search yielded 143 articles. After removing duplicates, non-peer-reviewed texts, or articles not explicitly related to modern agriculture, 89 papers were included in the analysis. The relevance and scope of threats were discussed, as well as empirical value and relevance to agricultural systems. A PRISMA-style diagram was not included but may be added in future work.

During the search, the most important criteria were:

- Does the study concern the use of IoT, AI, or cloud technologies in an agricultural context?
- Does the study address the threats, vulnerabilities, or risks associated with these technologies?
- Are the effects on the functioning or resilience of the food supply chain being analyzed?
- Is it an empirical study, a case study, a systematic review, or a meta-analysis of real implementations or incidents?
- Does the safety analysis specifically refer to applications in agriculture, rather than just technology in general?
- Does the publication include an analysis of security aspects, rather than just a description of technology implementations?

Fifty articles were found that meet all the criteria, and based on them, it can be concluded that agricultural systems utilizing IoT, AI, and cloud technologies are exposed to specific cybersecurity vulnerabilities that undermine the stability of the food supply chain. IoT devices suffer from weak authentication, lack of standardization, and network-level vulnerabilities, which enable unauthorized access and data leaks [5, 13, 17, 20]. AI systems are susceptible to data poisoning and model errors, which can lead to inaccurate predictions and erroneous decisions [1]. Cloud applications are susceptible to improper use of functions, which can result in data leaks and unauthorized access [16, 18], while blockchain

integrations are associated with vulnerabilities in smart contracts that can lead to financial losses [6].

Research indicates that these technical gaps affect various stages of the supply chain. Data breaches undermine information integrity and competitive position at all levels; network disruptions delay distribution and hinder inventory control; IoT device takeovers during production and processing stages cause equipment failures and quality issues; while ransomware attacks or supply chain-targeted attacks can disrupt organizational operations and lead to economic losses [9, 22]. Additional challenges arise from the geographical dispersion of agriculture, the seasonality of activities, and the reliance on legacy systems [2, 3]. Among the recommended actions are: improving device security, enhanced authentication and encryption protocols, quality assurance frameworks for AI systems, and tailored risk management practices aimed at protecting and stabilizing the food supply chain.

2 Technologies and Typical Threats

In the era of Agriculture 4.0 and 5.0, the widespread use of Internet of Things (IoT) devices, autonomous machines, and cloud-based analytical systems increases the agricultural sector's vulnerability to various forms of cyberattacks. Such attacks, known from other sectors of the economy, are becoming increasingly significant in the context of food security and the functioning of farms.

2.1 Attacks on IoT Devices

The use of IoT devices in agricultural systems is becoming standard. At the same time, they often have serious security vulnerabilities resulting from the very design of these devices (a simple sensor will not be able to implement complex protocols), leaving default passwords, lack of communication encryption, lack of standardization, or using simple, easily attackable data transmission protocols [1, 13, 20]. Example attack scenarios include taking control of: soil sensors, temperature monitoring systems, irrigation automation, which can lead to deliberate operational errors such as excessive watering or improper fertilization, or ultimately, due to small but consistent acts of sabotage, the emergence of unexpected crop failures. Unauthorized access to data can cause leaks and allow the Intruder to collect information that enables them to carry out an attack, not just on the IoT infrastructure itself, but also on the supply chain or other agricultural systems surrounding it. Additionally, infected devices can be used as part of a botnet to conduct DDoS attacks on other systems [5, 17, 22].

Hackers can intercept and modify data transmitted between IoT devices and farm management systems [7, 8]. Techniques such as Man-in-the-Middle (MitM) allow for the introduction of false data (e.g., incorrect soil moisture information), which can lead to incorrect management decisions, jeopardizing agricultural production.

IoT communication systems most often use wireless networks based on Zig-bee, LoRa, or Wi-Fi technologies and are therefore susceptible to interference, such as radio jamming, which can completely disconnect devices from central management systems [12]. The result is the suspension of the automation of processes such as irrigation or crop monitoring.

Autonomous agricultural machinery management systems (tractors, combines, robots) are often not regularly updated. For this reason, it is possible to inject malicious code into control systems, which can result in machine damage, crop destruction, or threats to human safety [11]. Autonomous agricultural vehicles can be targets of remote hijacking, which can lead to: malfunctioning in the fields, damage to infrastructure, and threats to human safety [14].

2.2 Attacks on Agricultural Data and Cloud Platforms

Modern farms collect and process large amounts of data. Local systems are not sufficient for this. Often, farmers, upon purchasing specific equipment, gain access to the ability to use cloud platforms. Unfortunately, often pleased with such simplification and the “offering” of such broad possibilities, they do not realize that they are giving away, often forever and without any control, their own data, globally very valuable, which service providers, unrestrained by law, use in various ways, and sometimes also do not guarantee appropriate security measures. Agricultural data (such as field maps, crop type maps, yield forecasts, soil condition information, sowing data, pest information, treatment methods, agricultural damage) are stored in the cloud and analyzed on platforms such as Climate FieldView or John Deere Operations Center. These data have high commercial value, are disseminated beyond the control of farmers, and are also targets of cyberattacks [9, 16, 18]. Possible attacks include:

- data theft and resale to competitors,
- sabotage through data modification (e.g., weather data),
- access blocking using ransomware and ransom demands.

If a country does not control access to data regarding agricultural infrastructure, the largest and most modern farms, and supply chains that other countries have, how can it guarantee food security for its citizens?

2.3 GPS Signal Jamming and Spoofing

Precision agriculture heavily relies on location data obtained through global navigation satellite systems (GNSS), including GPS. These systems are used for the automatic control of agricultural vehicles, conducting precise sowing and spraying, and monitoring soil and crop conditions with an accuracy of a few centimeters (RTK GPS).

Interference (jamming) or deliberate signal spoofing can cause deviations in machine navigation (e.g., incorrect travel paths), improper dosing of fertilizers or pesticides, physical damage to crops, such as the machine passing through

unprotected areas, and incorrect spatial data measurements for crop analysis [14, 22].

Research shows that GNSS signals are particularly susceptible to interference because they have very low power (they reach Earth from a considerable distance—around 20,000 km), and their transmission protocols were designed with open access in mind, not cybersecurity. GPS signal interference can be unintentional (e.g., electromagnetic interference) or intentional—using devices that disrupt GNSS systems (so-called GPS jammers). In the context of agriculture, this leads to a temporary loss of the ability to automatically control machines.

GPS spoofing is a more sophisticated attack that involves transmitting false GNSS signals that mimic correct positional data. As a result, the machine may be directed outside the proper work area, carry out agrotechnical activities in incorrect locations, and generate incorrect aggregate data about the field (which affects crop analyses and future production decisions).

GPS disruptions are regularly observed in Central and Eastern Europe, especially in the context of geopolitical tensions (e.g., along the eastern borders of the EU). In 2022, GPS problems were reported in the border regions of Poland and Lithuania, which translated into difficulties in the automatic execution of agricultural work in large farms.

2.4 Threats from AI Systems

With the rise in popularity of AI models worldwide, it was natural to harness them in agriculture. Artificial intelligence models are used for weather forecasting, crop monitoring, yield prediction, irrigation management, sowing, plant disease detection, and commonly understood advisory services. At the same time, the implementation of AI in agriculture requires an analysis of cybersecurity threats that could lead to serious disruptions in agricultural production and the food supply chain [10].

One of the threats when integrating AI systems with agricultural support systems is data poisoning. Attacks of this type are based on the deliberate introduction of malicious data into training sets, ultimately leading to incorrect decisions by AI systems. Such inappropriate advice coming from an AI model that has already been trusted can result, for example, in improper irrigation and losses in the fields. Research shows that such attacks reduce the performance of AI models and undermine trust in modern technologies.

Another problem that may arise when using AI models is backdoor attacks. In this situation, the AI model contains hidden mechanisms that are activated when the system meets certain conditions or through external activation. They lead to incorrect or even harmful system behavior. For example, one could ignore information about disease or lack of irrigation and suffer real crop losses. The most difficult part is that such backdoors can remain undetected until the moment of the attack.

Errors in the operation of systems coupled with AI models can arise from the inherent imperfections of the model itself. For example, through overfitting

or incorrect assumptions about the input data. In agriculture, it is often difficult to obtain complete data; we have many criteria to consider, or they are simply unpredictable. Model errors could lead, for example, to improper management of resources, pesticides, and fertilizers, which will affect the quantity and quality of the obtained yields. Model errors could lead to improper management of resources, pesticides, and fertilizers, which will affect the quantity and quality of the obtained yields.

Federated learning models are increasingly being used in agriculture, allowing AI models to be trained on distributed data without the need for centralization. Although this approach increases data privacy, it also introduces new opportunities for attacks. An intruder could, for example, poison the model by sending malicious updates from individual end devices, leading to errors in the model or the entire agricultural system [21].

3 Legal Regulations Concerning Agricultural Data

Every information system, including those supporting agriculture, is connected to data and its management. The article has already presented the cyber threat associated with them. However, there is a more elusive matter that poses an indirect or direct threat to food security. Around the world, there are various standards, laws, and models regarding data ownership and what can be done with it. Sometimes, even if the law is favorable to the data generator, there are gaps or issues with its enforcement. Here is an analysis of the situation based on several countries.

In the case of the European Union, the law seems to be quite restrictive and farmer-friendly. According to the law, the farmer or the entity generating the data is its owner. Data is protected in accordance with the General Data Protection Regulation (GDPR), meaning it should be shared only in necessary amounts, legally, based on contracts, and transparently. Data sharing must be based on clear agreements that specify the purpose, scope, and duration of data processing.

In the case of the United States, there are no uniform federal regulations. Each state has its own solutions in this matter. Often, this is left to the agreement between the farmer and the service provider. Some states, like Nebraska, are working on introducing their own regulations to protect agricultural data privacy, defining, among other things, the concepts of “agricultural data” and “data processing”. The American Farm Bureau Federation has developed principles regarding data privacy and security aimed at protecting farmers’ interests in their relationships with technology providers; however, these are not legally binding but rather good practices.

In Canada, agricultural data can be protected under the Personal Information Protection and Electronic Documents Act (PIPEDA) [27], provided that it contains personal data. Individual data ownership issues are resolved through contracts, hence the lack of top-down protection and regulations. A similar situation occurs in Australia [28]. There are general data protection regulations, and specific agricultural data issues are resolved at the contract level.

In India, the most contentious law is the one introduced in 2021 concerning geospatial data, which allows private entities access to spatial data without prior consent from the government or the landowner [29]. There is also a lack of regulations regarding personal data protection.

Since not every farm/agricultural enterprise in the world has access to legal advice, the issue of shifting regulations onto contracts is detrimental to farmers. Large corporations take data and use it with impunity, or worse, trade it with the unaware consent of farmers [19]. There is also the danger of a country losing critical data, as information about the structure of agriculture, its problems, and successes leaks out, where stronger countries can influence the decisions and agricultural resources of weaker economies.

The implementation of new legal solutions related to specific sectors of the economy, including agriculture, is often met with opposition from large corporations producing agricultural technologies. Lobbyists refer to the confidentiality of business data and numerous legal differences around the world, claiming that these are barriers that prevent the introduction of uniform regulations regarding the ownership of agricultural data.

4 Techniques for Defending Against Cyberattacks

Agriculture that relies on digital technologies is undoubtedly exposed to cyber threats. There is no universal security strategy, and effective protection must be “tailored” and based on a multi-layered approach that combines technical, organizational, and legal-formal security measures. Below, in Table 1, a summary of areas that may be subject to attacks is presented, along with their security methods and the protections or good practices applied in individual cases.

Despite the entire range of possible cyberattacks, what is most difficult remains outside the technical realm. Farmers should be regularly trained in cybersecurity because often the weakest link is the human being. The example of the aforementioned ransomware attack on NEW Cooperative Inc. stemmed, among other things, from the fact that most of the company’s employees used the same password: “chicken1”... Awareness of the importance of implementing appropriate password policies, monitoring equipment, and using proper barriers and techniques is crucial. The second difficult issue to grasp is legislation, which in most parts of the world does not side with farmers. Countries should finally recognize agricultural data as critical for their own security as well, introduce rules for signing contracts regarding this type of data, and safeguard them with principles similar to those for personal data or trade secrets.

Table 1. Defence Techniques

Security area	Method	Description
Security of IoT devices	Authentication and authorization	Enforcing strong authentication (including two-factor authentication) for all IoT devices used in farms [13].
	Network isolation	Dividing the network through the use of VLANs, firewalls, and dedicated gateways that restrict access exclusively in an authorized manner to specific devices [15].
	Communication encryption	Implementation of end-to-end encryption for data transmission between sensors and the cloud to prevent eavesdropping or tampering [12].
	Updates and vulnerability management	Regular firmware updates for IoT devices and the use of vulnerability management systems [18].
Defense against specific types of attacks	Jamming attacks	Introduction of distributed communication systems and low-power sleep mode in case of signal loss [12]
	Man-in-the-Middle (MitM) attacks	Dynamic reconfiguration of sensor networks through the creation of clusters with periodically changing communication and authentication methods [7, 8].
	GPS Interference	Integration of object recognition algorithms in the field and the use of physical markers for position verification [14].
	Security of auto- nomous machines	Implementation of IT standards for autonomous vehicle management systems, including testing resilience to control takeovers [14].
Protection of agricultural data and cloud platforms	Cloud data security	The use of advanced data encryption at rest and in transit, as well as multi-level access management [16].
	Data ownership protection	Implementing data policies that guarantee farmers retain ownership rights to the agricultural data generated, in accordance with the EU Code on the sharing of agricultural data [24].
	Data risk management	Regular assessments of the risk associated with data loss or unauthorized disclosure within the framework of information security management policies.
	Detection of backdoor attacks	Implementation of model auditing techniques and methods for detecting hidden mechanisms controlling model behavior [10].
Artificial intelligence systems	Explainable Artif. Intelligence (XAI)	The use of transparent models or XAI tools to increase the transparency of predictive systems [21].
	Securing federated learning	Implementing systems for detecting and isolating malicious participants in federated learning, e.g., through outlier-resistant models [21].
	AI systems security audit	Conducting regular audits and penetration tests of AI models used in farms.

5 Conclusions and Directions for Further Research

Agriculture is increasingly utilizing modern technologies and digitization, which should bring significant benefits to societies in terms of higher yields, reduced labor burden, optimized production, precise sowing/fertilization/irrigation, and increased efficiency. However, progress is accompanied by cybersecurity threats that can directly or indirectly affect the food security of societies. The analysis conducted in this article indicates that agricultural systems utilizing IoT technologies, artificial intelligence, autonomous machines, and cloud solutions are exposed to a range of serious risks. Among them, one should mention, among others: weak security of IoT devices, the possibility of taking control over autonomous machines, disruptions and falsification of GPS signals, the vulnerability of cloud systems to data leaks, and threats arising from imperfections or malicious manipulation of AI systems. A significant problem also remains the lack of coherent, unambiguous legal regulations regarding the ownership and protection of agricultural data, which exposes agricultural producers to losing control over one of the most valuable resources of the digital economy.

Let us also pay attention to the stakeholders responsible for individual areas of agricultural data security. First, governments - must create laws protecting agricultural data, because it is part of critical infrastructure. They should support safe technological solutions at the national level. Second, companies providing technological solutions - should ensure transparency of contracts and data management, ensure their security and process them in an ethical manner, which of course may remain our naive wish if there is no law enforcing this on entrepreneurs. Third, farmers - taking care of education, also in the matter of digital security is of great importance here. Farmers must have good habits in the form of a correct password policy, updates, access control, or device protection. Finally, international organizations - should coordinate all activities, proposing common technical, legal and ethical standards. Only joint action of these groups will increase the resilience of agriculture to cyber threats and secure global food supply chains.

In the context of further research, it is advisable to deepen analyses regarding the resilience of agricultural systems to GPS signal disruptions and to develop alternative positioning technologies. There is also a need to develop effective mechanisms for detecting and countering attacks on federated learning systems in agriculture. An important direction for development is also the implementation of explainable artificial intelligence techniques, which will enable users to better understand the functioning of predictive and recommendation systems, thereby increasing trust in new technologies.

Further empirical research is also needed to model the impact of cyberattacks on food supply chains, particularly in the context of logistical disruptions and the destabilization of agricultural markets. The need to create effective educational programs aimed at farmers and agricultural technology operators, which will increase their awareness of threats and their ability to respond to security incidents, cannot be overlooked.

The current study is conceptual, based on a synthesis of the literature, and aims to identify the most important elements for further research. Future analyses should focus on empirical verification of the listed threats and propose a practical framework for safety in real modern farms.

References

1. Alahe, M.A., et al.: Cyber security in smart agriculture: threat types, current status, and future trends. *Comput. Electron. Agric.* **216** (2024)
2. Alatalo, J., Sipola, T., Kokkonen, T.: Food supply chain cyber threats: a scoping review. In: *Proceedings of the WorldCIST'23 – World Conference on Information Systems and Technologies*, pp. 397–408. Springer, Cham (2023)
3. Alsinglawi, B., Zheng, L., Kabir, M.A., Islam, M.Z., Swain, D., Swain, W.: internet of things and microservices in supply chain: cybersecurity challenges, and research opportunities. In: *Advanced Information Networking and Applications. AINA 2022, Lecture Notes in Networks and Systems*, vol. 451, pp. 667–679. Springer, Cham (2022)
4. Berry, E.M., et al.: Food security and sustainability: can one exist without the other? *Public Health Nutr.* **18**(13), 2293–2302 (2015)
5. Bertino, E., Islam, N.: Botnets and internet of things security. *Computer* **50**(2), 76–79 (2017)
6. Bhat, S.A., et al.: Agriculture-food supply chain management based on blockchain and IoT: a narrative on enterprise blockchain interoperability. *Agriculture (MDPI)* (2021)
7. Cekerevac, Z., Dvorak, Z., Prigoda, L., Cekerevac, P.: Internet of things and the man-in-the-middle attacks – Security and economic risks. *MEST J.* **5**(2) (2017)
8. Conti, M., Dragoni, N., Lesyk, V.: A survey of man in the middle attacks. *IEEE Commun. Surv. Tutor.* **18**(3), 2027–2051 (2016)
9. Etemadi, N., Borbon-Galvez, Y., Strozzi, F.: Blockchain technology for cybersecurity applications in the food supply chain: a systematic literature review. *Comput. Electron. Agric.* **179** (2020)
10. Gao, Y., Li, Y., Gong, X., Shi, Q., Yuan, D., Zhu, L.: Security threats to agricultural artificial intelligence: position and perspective. *Comput. Electron. Agric.* **216**, 107086 (2024)
11. Goel, S.: A systematic literature review on past attack analysis on industrial control systems. *Trans. Emerg. Telecommun. Technol.* **35**(6), e5004 (2024)
12. Ingham, M., Marchang, J., Bhowmik, D.: IoT security vulnerabilities and predictive signal jamming attack analysis in LoRaWAN. *IET Inf. Secur.* **14**(4), 368–379 (2020)
13. Meneghello, F., et al.: IoT: Internet of threats? A survey of practical security vulnerabilities in real IoT devices. *IEEE Internet Things J.* **6**(5) (2019)
14. Parkinson, S., Ward, P., Wilson, K., Miller, J.: Cyber threats facing autonomous and connected vehicles: future challenges. *IEEE Trans. Intell. Transp. Syst.* **18**(11), 2898–2915 (2017)
15. Pacheco, J., Hariri, S.: IoT security framework for smart cyber infrastructures. In: *Proceedings of IEEE Foundations and Applications of Self-Systems Workshops (FAS-W)*. IEEE (2016)
16. Singh, J., Pasquier, T., Bacon, J., Ko, H., Eyers, D.: Twenty security considerations for cloud-supported internet of things. *IEEE Internet Things J.* **3**(3), 269–284 (2015)

17. Szymoniak, S., Piątkowski, J., Kurkowski, M.: Defense and security mechanisms in the internet of things: a review. *Appl. Sci.* **15**(2), 499 (2025)
18. Tejaswi, B., Mannan, M., Youssef, A.: Security weaknesses in IoT management platforms. *IEEE Internet Things J.* **11**(1), 1572–1588 (2023)
19. Wiseman, L., Sanderson, J.: *Farm Data and Privacy: What Farmers Need to Know*. Griffith University Research Repository (2020)
20. Yang, Y., Wu, L., Yin, G., Li, L., Zhao, H.: A survey on security and privacy issues in internet-of-things. *IEEE Internet Things J.* **4**(5), 1250–1258 (2017)
21. Yaacoub, J.-P.A., Noura, H.N., Salman, O., Chehab, A.: Security of federated learning with IoT systems: issues, limitations, challenges, and solutions. *Internet Things* **19**, 100565 (2022)
22. Yazdinejad, A., et al.: A review on security of smart farming and precision agriculture: security aspects, attacks, threats and countermeasures. *Appl. Sci.* **11**(16), 7518 (2021)
23. New Cooperative Ransomware Attack Timeline: Status Updates. <https://www.msspalert.com/news/new-cooperative-ransomware-attack-timeline-status-updates>. Accessed 29 Jan 2025
24. EU Agricultural Data Stakeholders Task Force, EU Code of Conduct on Agricultural Data Sharing via Contractual Agreements (2020). https://fefac.eu/wp-content/uploads/2020/07/eu_code_of_conduct_on_agricultural_data_sharing-1.pdf. Accessed 28 Apr 2025
25. Nebraska Legislature, LB525 - Adopt the Agricultural Data Privacy Act (2025). <https://www.billtrack50.com/billdetail/1797228>. Accessed 28 April 2025
26. American Farm Bureau Federation, Privacy and Security Principles for Farm Data (2014). <https://www.agri-pulse.com/ext/resources/pdfs/p/r/i/t/a/PrivacyAndSecurityPrinciplesForFarmData.pdf>. Accessed 28 Apr 2025
27. Government of Canada, Personal Information Protection and Electronic Documents Act (PIPEDA) (2000). <https://laws-lois.justice.gc.ca/eng/acts/p-8.6/>. Accessed 28 Apr 2025
28. Office of the Australian Information Commissioner, Privacy Act 1988 (1988). <https://www.oaic.gov.au/privacy/privacy-legislation/the-privacy-act>. Accessed 28 Apr 2025
29. Sinha, P.: India's Geospatial Data Revolution Raises New Surveillance Fears. *WIRED* (2021). <https://www.wired.com/story/india-geospatial-data-surveillance-business/>. Accessed 28 Apr 2025
30. Paganini, P.: Hackers hit Swiss farm and shut down cow-milking robot. *Security Affairs* (2023). <https://securityaffairs.com/166839/cyber-crime/cow-milking-robot-hacked.html>. Accessed 10 June 2025



Cybersecurity 2030: The Synergy Between Machine Learning and Generative Artificial Intelligence

Mário Marques da Silva^{1,2,3} , Fernando Correia^{1,2,3} ,
Héctor Dave Orrillo Ascama^{1,2,3} , Laercio Cruvinel Júnior^{1,2,3} ,
Guilherme Lopes Fernandes¹ , Tomás Viana¹ , and Afonso Frasilho¹ 

¹ Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa,
1169-023 Lisbon, Portugal

{mmsilva, fjcorreia, horrillo, lcruvinel}@autonoma.pt, {30010398,
30010623, 30010929}@students.ual.pt

² Autónoma TechLab, 1169-023 Lisbon, Portugal

³ Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal

Abstract. The rapid evolution of technology, characterized by the 4th Industrial Revolution, has reshaped the cybersecurity landscape. This paper explores the implementation of Generative Artificial Intelligence in the context of cybersecurity, highlighting their applications in data analysis procedures and automatic control systems. By integrating Machine Learning for real-time detection and Generative Artificial Intelligence for simulating advanced attack scenarios, we can detect cyberattacks at an early stage and minimize impact on system functionality. We conclude by discussing the implications for the future of cybersecurity and the anticipated dominance of AI-driven solutions by 2030.

Keywords: Cybersecurity and Supervised Learning · Generative Artificial Intelligence · Machine Learning · Hybrid Defense Strategy · Anomaly Detection · Synthetic Data

1 Introduction

The proliferation of digital systems in the 4th Industrial Revolution has introduced unprecedented cybersecurity challenges. This paper explores the implementation of Machine Learning (ML) and Generative Artificial Intelligence (GEN AI) in the context of cybersecurity, highlighting their applications in data analysis procedures and automatic control systems. Integrating ML for real-time detection with generative-AI simulations enables advanced attack scenarios, we can detect cyberattacks at an early stage and minimize impact on system functionality. We conclude by discussing the implications for the future of cybersecurity and the anticipated dominance of AI-driven solutions by 2030. With the advent of interconnected devices and the Internet of Things (IoT), cyber-attacks have become more sophisticated, necessitating advanced defence

mechanisms. Artificial Intelligence (AI) has emerged as a cornerstone in this domain, offering tools for automation, real-time threat detection, and strategic simulation [1]. This paper examines the use ML methodologies, which rely on labelled datasets [2], and GEN AI, which leverages unstructured data to simulate and preempt cyber threats [3].

This paper is organized as follows: Sect. 2 relies on the background and literature review; Sect. 3 presents the methodology used in this paper; Sect. 4 shows the simulation results and discussion, while Sect. 5 concludes this paper.

2 Background and Literature Review

According to Capodieci et al. [4], there are different types of AI, such as Machine Learning (ML), and Deep Learning (DL), and GenAI. While ML depends on the human factor to classify and contextualize data, DL is a system with greater autonomy in processing unstructured data. Also, Large Language Models (LLM) represent another type of AI technology, based on GenAI, capable of understanding and generating written language [5].

In cybersecurity, given the volume of data that needs to be analysed to identify possible threats, the human factor can be a limiting factor. However, ML has been a fundamental technique in cybersecurity, excelling in detecting known attack patterns and automating routine tasks such as log analysis [6], and to do that, human interaction is required.

DL, an advanced subset of ML, enhances these capabilities through multilayer neural networks [7]. DL is part of unsupervised learning, which is effective in anomaly detection and clustering, identifying previously unseen threats without the need for labeled data.

GEN AI, powered by large language and multimedia models, extends these capabilities by creating synthetic data, simulating attack scenarios, and even mimicking social engineering tactics [3].

However, its potential misuse, such as generating deepfakes or creating malware variants, highlights the dual-edged nature of this technology [8]. According to the 2024 report by Capgemini [9], in the universe of 1000 organizations analysed, around 45% have encountered deepfake attacks in the last 2 years, and in 43% there have been financial losses due to the exploitation of deepfake technology.

GEN AI operates along two main vectors: one focused on attacks and the other on testing and strengthening defences. In the attack vector, as mentioned earlier, GEN AI is used to create new forms of attacks on systems by exploiting unknown vulnerabilities. On the other hand, the same vector can be proactively applied to generate simulated attack scenarios, which enables the training and continuous improvement of detection systems. This duality not only expands the understanding of emerging threats but also strengthens cyber defences by anticipating and mitigating potential future attacks. In this way, it is possible to leverage the capabilities of ML in combination with GEN AI. While ML can learn to detect known attacks, GEN AI uses this acquired knowledge to infer patterns and generate simulated attack scenarios. This combined approach contributes to minimizing the impact of AI-driven cyberattacks [10].

3 Methodology

To explore the integration of ML and GEN AI in cybersecurity, we analyzed:

- Applications of ML, including classification, regression, and anomaly detection. While Supervised Learning (SL) clearly identifies an attack type, Unsupervised Learning (UL) can be used to detect an event that is outside of the normal activity.
- GEN AI can be applied in cybersecurity to simulate phishing attacks, malware behaviors, and potential intrusion scenarios, enhancing the ability to predict and mitigate cyber threats.
- A hybrid defense strategy combining the strengths of both paradigms.

SL can be used as a classifier, where labels are assigned to data to identify the categories to which that data belongs. For example, the classification of animal images by identifying which animal each image belongs to. It can also be used in regression when it involves the continuous prediction of numerical values based on the input data. Finally, object detection which focuses on identifying and locating multiple objects in images or video. Examples of SL can be found at car park license plate identification processes for automatic payment systems, weather forecasting, credit card fraud detection, among others. On the other hand, UL can be employed to identify anomalous patterns in real-time events. This normally requires the analysis of an operator, to identify or not as a threat.

ML makes it possible to collect data and produce data resulting from previous experiences, improving information quality. In other words, with increased knowledge, performance criteria can be optimized. ML makes it easier to solve various computational problems involving classification and regression procedures. Generically, ML allows for better control over data classes in AI training.

Currently, TBytes of data are produced, so classifying them using human supervision is challenging. Training an ML model requires a lot of computing time and supervision time. ML cannot perform all the complex tasks in the field of Machine Learning, which can be a disadvantage in the case of applying this model. The employment of ML and GEN AI must be weighed up and applied according to the volume of data to be processed and the results to be obtained.

Thus, a hybrid security system can be established, integrating not only classification, regression, and anomaly detection processes but also proactive mechanisms. These mechanisms enable the simulation of attack scenarios generated by GEN AI, which can be used to retrain and enhance the ML detection system, improving its ability to identify and respond to emerging threats.

Our study involved simulated environments using real-world datasets and generative models trained to mimic cyberattack scenarios. Metrics for evaluation included detection accuracy, response time, and the ability to pre-emptively counter novel threats.

4 Simulation Studies

This section aims to demonstrate that supervised models like Random Forest are effective in classifying multiple attack classes. Unsupervised models like Autoencoder are effective in detecting anomalous patterns, complementing the supervised models. The

proposed hybrid approach improves detection accuracy and reduces response times, enabling organizations to mitigate threats before they escalate.

The use of ML for intrusion detection (Intrusion Detection Systems [IDS]) has become prominent in the field of cybersecurity, enabling the analysis of large volumes of network traffic and the identification of anomalous or malicious behavior. This approach facilitates early attack detection and real-time automated monitoring.

In this context, it is pertinent to consider how IDS can benefit from the integration of multiple ML approaches. The use of a hybrid system, combining supervised and unsupervised models, emerges as an effective response to the growing sophistication of cyberattacks. Supervised models excel in accurately detecting known patterns, thanks to training with previously labeled data. However, they reveal limitations when faced with novel attacks or variants of existing threats. It is in such situations that unsupervised models play a crucial role, allowing the identification of anomalous patterns in real time, even without prior knowledge, acting as additional sensors for unexpected behaviors within the network.

4.1 Characterization of the Simulation Scenario

A complementary hybrid approach is illustrated in Fig. 1, combining the accuracy of the supervised model with the adaptive capabilities of the unsupervised model. Initially, network traffic is analyzed by a supervised model trained to recognize patterns associated with previously known attacks, immediately generating an alert whenever a threat is identified. If the supervised model does not detect any threat, the traffic is then analyzed by an unsupervised model, which clusters network behavior into similar patterns. Significant deviations from the established patterns may indicate anomalous behaviors, enabling the detection of unknown or newly emerged attacks. When an anomaly is detected, the probability of the traffic being malicious is estimated, and only the cases that exceed a defined threshold are forwarded for manual verification, thereby reducing the occurrence of false positives and minimizing the manual analysis workload.

Having introduced the concept of a hybrid approach to intrusion detection, it becomes necessary to select and prepare appropriate data to support its development and future validation. In this context, the present study is based on the “CSE-CIC-IDS2018” dataset, developed by the Canadian Institute for Cybersecurity (CIC) in collaboration with the Communications Security Establishment (CSE) of Canada for Cybersecurity [11]. The dataset was generated from real network traffic captured in a controlled environment, encompassing both benign communications and a variety of attack types (brute-force, DoS, DDoS, botnets, SQL injections), distributed across several CSV files, each corresponding to a day of simulation.

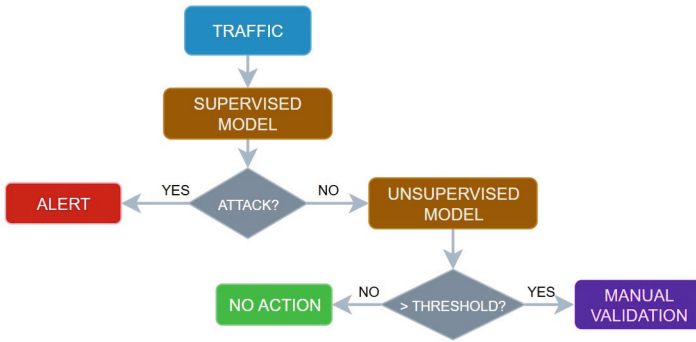


Fig. 1. Hybrid IDS operation flow

4.2 Results

A consolidated version was constructed using data from six specific days (February 14, 15, 21, 22, 23, and March 2, 2018), totaling 162,658 records and 78 numerical features. The integration and preprocessing steps were conducted using Python scripts, leveraging libraries such as “pandas” for data loading, transformation, and unification.

Among the available features, several stand out for their relevance in characterizing network traffic:

- Flow Duration: total flow duration in microseconds;
- Tot Fwd Pkts/Tot Bwd Pkts: total number of packets sent in the forward and backward directions;
- TotLen Fwd Pkts/TotLen Bwd Pkts: total packet length (in bytes) in both directions;
- Fwd Pkt Len Mean/Bwd Pkt Len Mean: average packet length, essential for understanding client-server communication patterns;
- Flow Bwts/s/Flow Pkts/s: average transmission rate in bytes and packets per second;
- Flow IAT Mean/Std/Max/Min: inter-arrival times between successive packets within a flow, useful for identifying irregular temporal patterns;
- Pkt Len Min/Max/Mean/Std/Var: descriptive statistics of packet length, useful for detecting uncommon or suspicious flows;
- SYN Flag Cnt / ACK Flag Cnt/FIN Flag Cnt: TCP flag counters, important for detecting typical attack behaviors such as SYN Floods;
- Idle Mean/Max/Min: idle time between communications, which may indicate evasion attempts.

All features were standardized into numerical formats (int64 and float64), ensuring consistency across files.

The statistical analysis revealed several noteworthy patterns. Flow Duration shows a mean of 105,457 μ s and a median of 37,634 μ s, with strong skewness (−236.88) and high kurtosis (59,458.29), indicating the presence of outliers. Tot Fwd Pkts displays negative minimum values and a skewness of −2.05. Meanwhile, Tot Bwd Pkts shows a skewness of 82.61, suggesting bursts of intensive traffic. These calculations were obtained using the “describe()” function and additional statistical tools from the “NumPy” library. These characteristics emphasize the need for normalization and rigorous data validation.

The class distribution is illustrated in Fig. 2. The “Benign” class includes 100,000 records (61.48%). Classes 2 to 7, with 10,000 records each (6.15%), represent the primary attack types. Classes 8 to 11 are less represented: 1,730 (1.06%), 611 (0.38%), 230 (0.14%), and 87 (0.05%) records, respectively.

The “Label” variable identifies each type of traffic:

1. benign;
2. FTP-BruteForce;
3. SSH-Bruteforce;
4. DDOS attack-HOIC;
5. Bot;
6. DoS attacks-GoldenEye;
7. DoS attacks-Slowloris;
8. DDOS attack-LOIC-UDP;
9. Brute Force-Web;
10. Brute Force-XSS;
11. SQL Injection.

This multiclass structure enables the training of models capable of distinguishing between normal traffic and multiple attack patterns. The intentional imbalance reflects real-world network conditions, where benign traffic predominates. This favours model training by reinforcing the recognition of normal patterns. The distribution was visualized using “matplotlib.pyplot.bar” on the values of the “Label” variable.

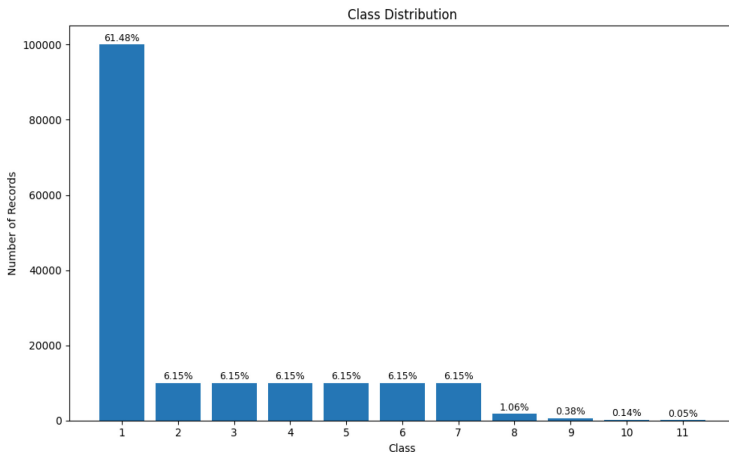


Fig. 2. Class distribution

Correlation analysis was conducted using the 15 features with the highest variance (Fig. 3), focusing on the most informative ones. Strong correlations were observed among time-based metrics related to inactivity, such as Idle Max, Idle Mean, and Idle Std (all near 1.00), as well as among inter-packet timing features like Fwd IAT Std with Flow IAT Std (0.94) and Fwd IAT Mean (0.94). The correlation matrix was generated with

“seaborn.heatmap” using the Pearson method on numerical features. These redundancies suggest overlapping information, and their removal or consolidation during feature selection is recommended to optimize model performance and reduce complexity.

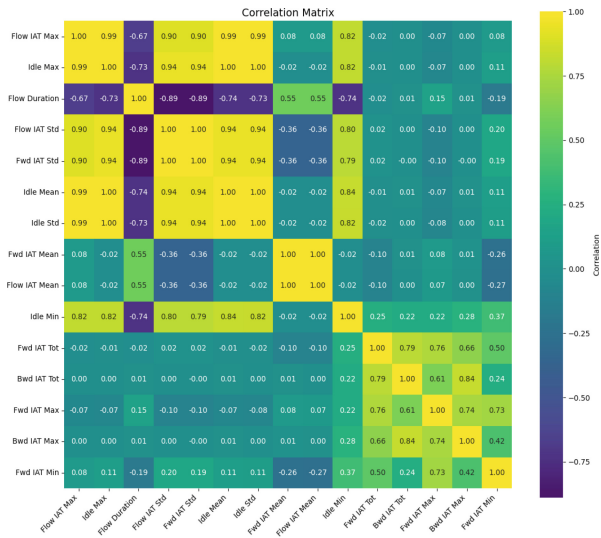


Fig. 3. Correlation matrix

4.3 Results Analysis

After defining the hybrid architecture and preparing the data, ML models were selected to ensure robustness and adaptability in intrusion detection. For the supervised component, three models were tested: Multilayer Perceptron (MLP), Random Forest, and Support Vector Machine (SVM). In the unsupervised domain, Isolation Forest, K-Means, and Autoencoder were analyzed.

The SVM model was included due to its effectiveness in separating classes in high-dimensional spaces. After normalization with StandardScaler and splitting the data into 70% for training and 30% for testing, it was trained using the SVC classifier. The evaluation on the test set revealed a weighted F1-score of 0.99, along with equally high values for precision and recall (both 0.99). The confusion matrix (Fig. 4) shows an almost perfect performance across the first seven main classes. However, the minority classes (9, 10, and 11) exhibited lower recall values: 68%, 45%, and 29%, respectively. Ten-fold cross-validation confirmed the model’s consistent performance, with an average of 98.50%.

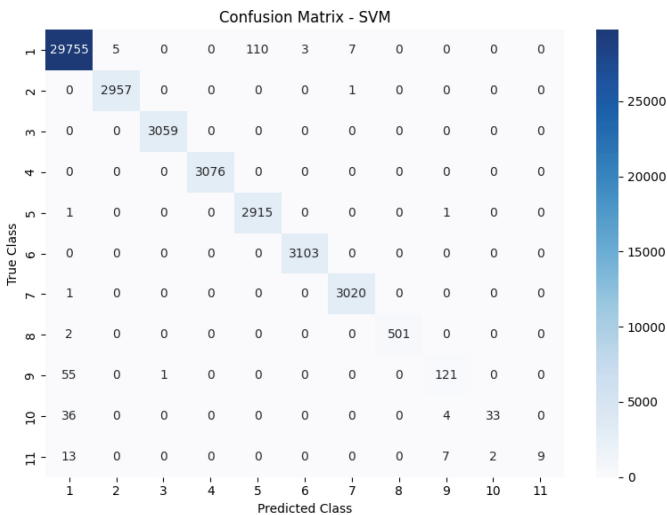


Fig. 4. SVM Confusion Matrix

Despite the model’s strong performance, its limitation in detecting rare attacks led to the exclusion of SVM as the final model. The Multilayer Perceptron (MLP) was selected for its ability to capture nonlinear patterns. It was trained with a single hidden layer of 100 neurons and 500 iterations, aiming to balance performance and computational cost. The evaluation showed weighted precision, recall, and F1-score of 1.00. The confusion matrix (Fig. 5) demonstrates strong results for the main classes and improvements over SVM in classes 9, 10, and 11, with recall values of 88%, 71%, and 29%, respectively.

Cross-validation confirmed the model’s stability ($99.12\% \pm 0.48\%$). However, increasing the number of neurons or layers could potentially improve the results, but at the cost of longer training times and higher computational resource consumption, which would not be justifiable compared to the next model.

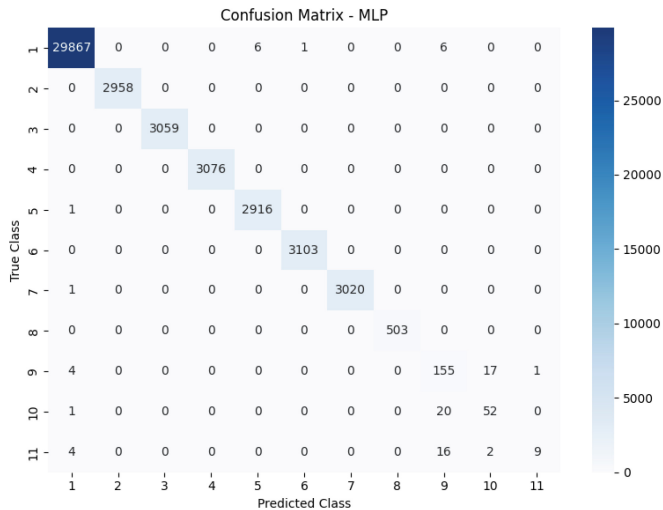


Fig. 5. MLP Confusion Matrix

The Random Forest model was included due to its effectiveness in multiclass classification and its generalization ability resulting from the combination of multiple decision trees. Using default parameters, it achieved a weighted F1-score of 1.00. The confusion matrix (Fig. 6) shows excellent performance in the less represented classes, with recall values of 92%, 79%, and 58% for classes 9, 10, and 11, respectively - outperforming both the MLP and SVM models.

Ten-fold cross-validation confirmed this robustness, with an average F1-score of 99.11% ± 0.57%. Considering the results obtained, Random Forest proved to be the most suitable choice for the supervised component, offering the best balance between accuracy, generalization capacity, and performance in detecting underrepresented classes, without compromising computational requirements.

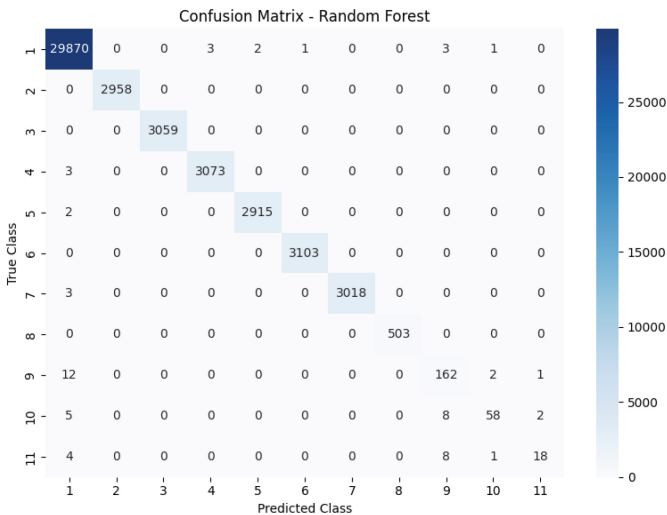


Fig. 6. Random Forest Confusion Matrix

The Isolation Forest model was used as a classical unsupervised approach for anomaly detection, leveraging its ability to isolate outliers through randomly constructed binary trees. The data were converted into a binary classification (0-benign traffic, 1-attack), and the model was trained using a contamination rate matching the proportion of attacks in the training set. Despite being computationally efficient, its performance was unsatisfactory due to its inability to capture the specific characteristics of the traffic in the dataset. The model assumed an anomaly distribution that did not reflect the real complexity of the data, flagging unusual but legitimate behaviours as attacks while ignoring subtle malicious patterns. On the test set, it achieved an F1-score of only 0.32, with equally low precision and recall, failing to reliably distinguish between normal and anomalous traffic. The confusion matrix (Fig. 7) shows a high number of false positives and false negatives, seriously limiting the model’s applicability in real-world scenarios.

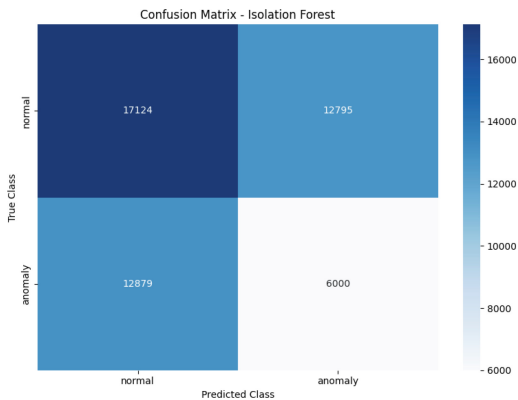


Fig. 7. Isolation Forest Confusion Matrix

Given these results, and despite the advantages of Isolation Forest in terms of simplicity and speed, its poor performance led to the rejection of this model.

The K-Means model was tested as an unsupervised approach for intrusion detection, based on the assumption that normal and malicious traffic would naturally form two distinct clusters. After converting the classes to binary (0- benign, 1 _ attack), the model was trained with two centroids. To align the output with the actual classification, it was necessary to invert the predictions so that the clusters correctly matched the “normal” and “anomaly” labels. The results were unsatisfactory. The model achieved an F1-score of only 0.43, with highly unbalanced performance across classes: the _normal_ class had a recall of just 26%, while the _anomaly_ class reached 60%. The confusion matrix (Fig. 8) highlights a significant number of false positives and false negatives, with over 22,000 benign samples incorrectly classified as anomalies and more than 7,500 attacks going undetected.

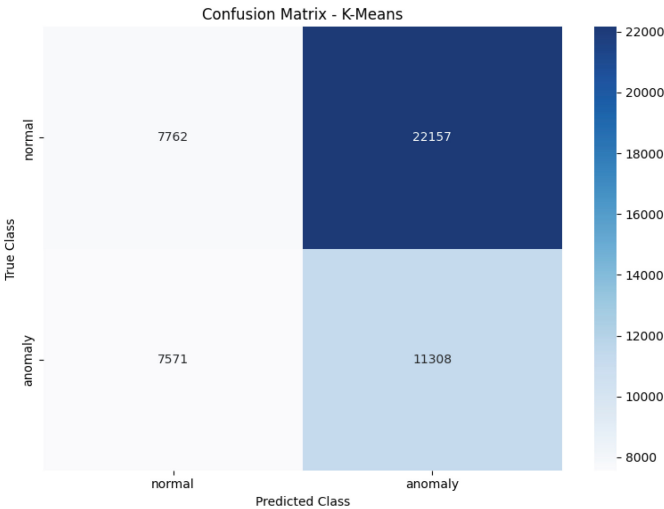


Fig. 8. K-means Confusion Matrix

These values demonstrate K-Means’ inability to capture the complexity and overlap of patterns present in network traffic. The algorithm’s linear nature - assuming spherical clusters with similar variance - proved incompatible with the actual structure of the data. This limitation justifies the rejection of K-Means as a model for the unsupervised component of the hybrid system.

The Autoencoder model was implemented with the goal of learning to reconstruct normal traffic patterns, if anomalies would yield significantly higher reconstruction errors. It was trained exclusively on benign data, using a symmetric neural network architecture with L1/L2 regularization and optimization via EarlyStopping. Evaluation was conducted on data containing both benign and malicious traffic, converted into a binary classification (0 – benign, 1 – attack).

The distinction was based on the mean squared reconstruction error (MSE), and the optimal threshold was determined using the precision-recall curve, selecting the point that maximized the F1-score. The resulting value was 0.000003, which may seem extremely low at first glance. However, this is justified using MinMaxScaler, which normalizes the data to the [0, 1] range, naturally reducing the scale of reconstruction errors. The results showed excellent anomaly detection capability: the model achieved an F1-score of 0.64, with a precision of 0.48 and recall of 1.00. The confusion matrix (Fig. 9) highlights 18,788 attacks correctly detected and only 91 false negatives, indicating high sensitivity to anomalous patterns. In contrast, 20,650 false positives were recorded, representing a considerable rate of incorrect alerts. Despite the high number of false alarms, the model’s sensitivity makes it particularly suitable as a complementary layer in a hybrid system, where it can act as an anomaly sensor for traffic that deviates from known patterns.

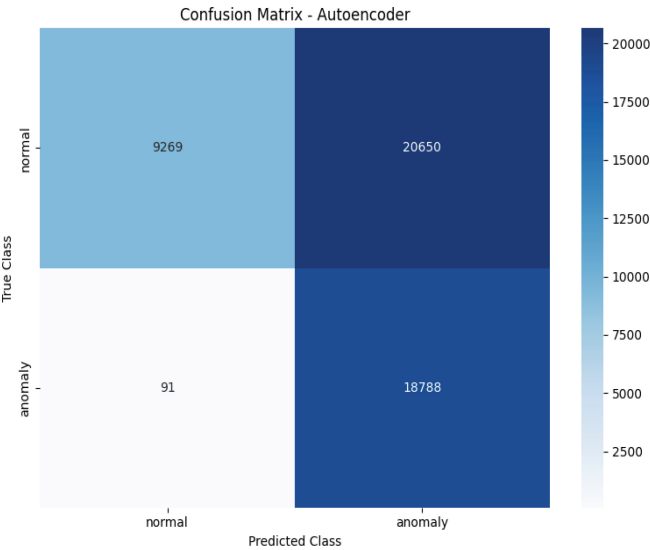


Fig. 9. Autoencoder Confusion Matrix

Although it does not reach the precision levels of the supervised models, the Autoencoder proved to be the most effective approach within the unsupervised domain. Its ability to adapt to new patterns and detect previously unknown traffic makes it an ideal complement to the supervised model, strengthening the hybrid system in scenarios involving uncatalogued attacks. Moreover, since it is trained exclusively on benign traffic, it can be updated periodically or even continuously with new unlabeled network data, allowing it to evolve in real time. It is also important to emphasize that all anomalies detected by this model are subject to manual validation, ensuring that potential false positives have no direct impact and preserving the overall reliability of the system.

5 Conclusions

The integration of ML and GEN AI presents a transformative approach to cybersecurity, offering a robust defense strategy against increasingly sophisticated threats. ML excels in real-time anomaly detection and classification of known attack patterns, while GEN AI enhances these capabilities by simulating novel attack scenarios and generating synthetic data for proactive defense training. This hybrid model not only improves detection accuracy but also reduces response times, enabling organizations to mitigate threats before they escalate. As cyberattacks grow in complexity, the synergy between these technologies will be critical in maintaining resilient and adaptive security systems.

However, the dual-edged nature of GEN AI cannot be overlooked. While it strengthens defenses, its potential misuse - such as creating deepfakes or advanced malware - poses significant risks. The findings of this study underscore the importance of ethical oversight and continuous human supervision to ensure these technologies are deployed responsibly. Organizations must balance innovation with caution, implementing safeguards to prevent malicious exploitation while leveraging GEN AI's potential to anticipate and neutralize emerging threats.

Looking ahead, the dominance of AI-driven solutions in cybersecurity is inevitable. By 2030, AI is expected to resolve up to 95% of cybersecurity incidents, revolutionizing how threats are identified and countered. To stay ahead in this arms race, institutions must invest in both ML and GEN AI, fostering collaboration between human expertise and automated systems. The future of cybersecurity lies in adaptive, intelligent frameworks that combine the precision of supervised learning with the creativity of generative models, ensuring a safer digital landscape for all.

References

1. Samuel, A.L.: Some studies in machine learning using the game of checkers. *IBM. J. Res. Dev.* **3**(3), 210–229 (1959)
2. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. *Neural Comput.* **18**(7), 1527–1554 (2006)
3. OpenAI. Gpt-4 technical report (2024)
4. Capodieci, N., Sanchez-Adames, C., Harris, J., Tatar, U.: The impact of generative AI and LLMs on the cybersecurity profession. In: 2024 Systems and Information Engineering Design Symposium (SIEDS), pp. 448–453 (2024)
5. Lee, A.: What are large language models and why are they important? <https://blogs.nvidia.com/blog/2023/01/26/what-are-large-language-models-used-for/>, January 2023
6. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning (2016)
7. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. The MIT Press (2016)
8. Brynjolfsson, E.: *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W.W. Norton & Company (2014)
9. Capgemini. New defenses, new threats: What AI and gen AI bring to cybersecurity. Technical report, Capgemini Research Institute (2024)
10. Al Siam, A., Hassan, M.M., Bhuiyan, T.: Artificial intelligence for cybersecurity: a state of the art. In: 2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC), pp. 1–7 (2025)
11. Canadian Institute for Cybersecurity. A realistic cyber defense dataset (csecic-ids2018) was accessed on apr2025 from <https://registry.opendata.aws/cse-cicids2018>. (2018)



Artificial Intelligence and Ethical Responsibility: The Impact of Algorithmic Decisions on Cybersecurity

Sabina Szymoniak¹(✉)  and Shalini Kesar²

¹ Department of Computer Science, Częstochowa University of Technology, 42-201 Częstochowa, Poland

sabina.szymoniak@icis.pcz.pl

² Department of Computer Science & Information Systems, Southern Utah University, Cedar City, UT 84720, USA

kesar@suu.edu

Abstract. The development of artificial intelligence creates new prospects in areas that include cybersecurity and raises critical ethical questions. When using AI in vital applications like safeguarding personal information or managing energy structures, it is crucial to assess the impact of wrong algorithms. Some examples include intrusion detection systems and intrusion prevention systems that employ machine learning models to monitor and analyze data in real-time to detect threats. The main issues are the system's opacity (the so-called 'black box' issue), the possibility of bias in the data used, and legal liabilities. The lack of transparency can lead to difficulties in understanding AI decisions, thus affecting user trust and allowing attackers to exploit algorithms. Bias in data can lead to unfair decisions and also tends to weaken security mechanisms. Also, it is still unclear who will be responsible for the AI failures, particularly regarding the data breaches and financial losses. To address these challenges, it is essential to promote the development of Explainable AI technologies, teach people how to use AI, check data, and adopt new legal requirements for AI. The paper analyses these problems and calls for a cross-disciplinary strategy encompassing technological, ethical, and legal perspectives to make the use of AI in cybersecurity more reliable.

Keywords: Artificial Intelligence · Cybersecurity · AI Ethics · Algorithm Transparency · Data Bias

1 Introduction

Artificial intelligence (AI) development has brought revolutionary changes to medical and educational fields as well as transportation systems and cybersecurity infrastructure. The increasing use of AI in critical systems creates multiple ethical and legal challenges regarding who should be responsible for decisions made by autonomous algorithms. As cybersecurity increasingly integrates AI-based solutions to detect threats and protect data infrastructures, the stakes associated with algorithmic decisions become ever higher [30, 31].

Cybersecurity systems today often depend on machine learning (ML) models and deep neural networks to monitor network activity, identify anomalies, and react to potential threats in real-time. The efficiency and scalability of AI come with new vulnerabilities that need to be addressed. The “black box” problem of algorithmic opacity and bias in training data and legal accountability for AI-driven decisions demonstrate the necessity of a multidisciplinary approach to responsible AI use [3, 7].

Errors or biases embedded in AI systems can have far-reaching consequences. Incorrect threat identification may compromise legitimate operations, erode user trust, and even expose organisations to significant legal liabilities. Moreover, the lack of transparency in AI decision-making complicates efforts to audit, control, and assign responsibility when failures occur [9, 10].

The research addresses multiple challenges by performing a complete evaluation of AI deployment risks in cybersecurity through ethical and technological and legal perspectives. The research presents a multi-dimensional framework which unites innovation with accountability. The research combines technical assessments with ethical considerations and legal frameworks by using recent scientific findings and EU Digital Responsibility Act regulations. The paper develops principles for AI protection of critical cyber infrastructures through its interdisciplinary method which ensures safety and transparency and trustworthiness [20, 25].

The rest of this paper is organized as follows. After an introduction that outlines the main issues related to AI and responsibility in the context of cybersecurity, the next part focuses on the opportunities and threats that AI poses in this area. Then, ethical and legal issues are discussed – both theoretically and practically. The next part proposes possible ways to implement such solutions responsibly. There is also room for two specific real-life examples that clearly demonstrate the complexity of the topics discussed. The whole ends with an attempt to organize the key conclusions and a reflection on where the development of AI in cybersecurity is heading.

2 AI in Cybersecurity: Opportunities and Risks

Artificial intelligence has expanded its usage in cybersecurity at a fast pace during the last few years through promising solutions while introducing fresh security dangers. AI systems now serve IT infrastructure protection and intrusion detection while monitoring network traffic and providing immediate responses to newly identified threats. Machine learning-based solutions combined with deep neural networks analyze large data sets at speeds surpassing traditional methods so they can identify faint abnormalities that human experts might miss. IDS and IPS systems with AI enhancement serve as examples of tools which monitor network traffic patterns continuously to detect suspicious behaviors before attacks can escalate [1, 6].

AI adoption in cybersecurity brings its most significant benefit through its learning capacity and adaptability features. AI models possess the ability to detect unanticipated attack patterns independently from predefined signatures because they operate unlike rule-based systems. Through its ability to adjust to new situations AI systems become effective at detecting zero-day attacks and emerging malware strains and sophisticated phishing attempts that traditional systems would miss. AI performs various security

operations automatically including incident classification and alert prioritization and initial response execution which enables human analysts to handle complicated cases [11, 28, 29].

The implementation of AI in cybersecurity systems provides multiple benefits but creates multiple new security issues. AI models function at their best when trained with data of high quality because any biases or gaps or inaccuracies in training datasets will negatively affect their performance. AI systems require training data that fully represents the variety of real-world threats and behaviors because inadequate representation leads to either missed threats or unnecessary false positives that waste resources [27, 33].

The term adversarial attacks describes an essential security risk. Threat actors use subtle methods to modify AI system inputs which results in incorrect classification of malicious behavior as well as false identification of benign activities. The phenomenon demonstrates a critical weakness in numerous machine learning models because they react badly to specific well-designed perturbations. Attackers who understand how AI-based security defenses function will likely try to exploit their weaknesses to remain undetected through security measures [2, 34].

The difficulty of understanding AI system operations emerges as the technology becomes increasingly sophisticated. Deep learning models operate as “black boxes” because their internal decision-making processes remain difficult for both creators and others to understand. The absence of explainable decision-making in cybersecurity systems becomes hazardous because it enables systems to block valid activities without proper explanation while missing actual threats without cause. Such issues reduce user trust levels and make incident auditing more complicated while creating legal risks because organisations cannot establish accountability for errors [12, 15, 32].

AI brings transformative capabilities to improve cybersecurity through speed, scalability and adaptability which human defenders cannot replicate on their own. The implementation of AI systems requires thorough examination of newly emerging risks such as data bias together with adversarial vulnerabilities and explainability and accountability issues. AI-based cybersecurity integration needs both technological development and ethical understanding and robust governance systems to guarantee secure and effective usage of powerful tools.

3 Ethical Challenges in AI Decision-Making

The integration of artificial intelligence into cybersecurity systems creates both promising possibilities and substantial ethical dilemmas. Advanced algorithms present a major problem because they lack transparency in their decision-making processes. Modern deep learning-based models function in ways that remain unclear even to their developers. When AI flags legal activity as a threat or – even worse – fails to detect a real attack, it is difficult to determine why the failure happened. This is a huge challenge for administrators and users: how can you trust a system whose decisions cannot be explained? [22, 26].

The lack of transparency is not just a technical problem. It also undermines trust in technology and the organisations that implement it. Trust is key in cybersecurity. If users begin to doubt the sense of security, its effectiveness will significantly decrease.

AI-based systems need to be designed with built-in capabilities to provide some level of understanding about their decision-making processes [24].

The learning data for AI models contains serious biases which represent another major challenge. AI systems operate without objectivity because they learn from the information we supply. The algorithm will begin reproducing errors and biases that exist in the provided data. The specific conditions in cybersecurity lead to false threat identifications of particular user groups based on their location and device type. Such unfair treatment reduces the effectiveness of protection systems and carries the risk of violating equality and non-discrimination regulations [5].

The problem of bias is more than just technical imperfections. It is a social threat. Systematically treating certain groups of users worse can deepen existing inequalities and lead to serious legal consequences for companies implementing such solutions. Therefore, already at the stage of training models, it is necessary to take care of the diversity of data and regularly monitor their results to detect undesirable patterns as early as possible [16].

The ethical challenge is the issue of liability for AI errors. When AI makes a mistake, who should be held responsible? Should we hold the creator of the algorithm responsible? Who implemented the algorithm? Could it be the company that implemented it, or perhaps the individual who utilised the pre-made system? The need to establish liability becomes essential in cybersecurity because careless decisions can result in data breaches and financial damage and severe privacy violations. The absence of clear rules puts companies and users at risk of dangerous legal consequences while denying victims of errors their right to pursue legal action [18].

Ethical issues related to AI decisions in cybersecurity are therefore not just the domain of engineers or lawyers. These are challenges that affect everyone: from system designers to users to society as a whole. If we want AI to support security rather than create new risks, we must ensure transparency, fairness, and accountability at the stage of its creation and implementation.

4 Legal Responsibility for Algorithmic Decisions

The implementation of artificial intelligence in cybersecurity creates multiple technical and ethical problems which lead to significant legal liability concerns. Traditional systems depend on human personnel including administrators engineers and operators to make decisions about data protection and threat response. The process of determining accountability for autonomous algorithm decisions becomes significantly more complex [35].

The responsibility for errors committed by AI-based systems extends to multiple entities. The creators of the algorithm including programmers and companies that designed the model should bear responsibility. The discussion about unintentional incidents becomes difficult when errors result from inadequate design or insufficient testing. The organization that implements the system bears responsibility for its operation when it fails to provide proper supervision and regular updates and monitoring of the AI system. There is also the question of the data providers on which the model was trained – if the data was biased, incomplete or misleading, this had a direct impact on the functioning of the entire system [21].

The situation is complicated by the fact that many AI models operate in a way that is difficult to trace and explain. In the case of classic IT systems, it is simple to point out an error in the code or in the decision-making process. The internal operations of deep learning models in AI systems generate decisions which their developers cannot always understand. The technical and legal challenges of assigning blame to a particular person or entity make this task both difficult and complicated [19].

Current legal regulations in many countries do not keep up with the reality created by AI. Traditional liability frameworks, based on the assumption of human action or a classic system, are insufficient in a world where decisions are made by a self-learning algorithm. That is why there is increasing talk about the need to introduce new regulations that will clearly define the principles of liability for decisions made by AI. The European Union is preparing legislation known as the Digital Responsibility Act, which is a significant step in this direction. Its aim is to create a legal framework that will take into account the specifics of autonomous systems and ensure that their creators, suppliers and users bear appropriate responsibility for the consequences of their actions [35].

In the context of cybersecurity, this issue is particularly important. The organisation must be certain who is responsible for the damage if a threat is incorrectly classified, leading to a data leak or a breach of user privacy. Without clear liability rules, both victims of errors and companies implementing AI themselves may find themselves in a difficult legal situation, and the financial and image consequences may be enormous [19, 21].

Therefore, when developing and implementing artificial intelligence in security systems, one cannot limit oneself to technical risk analysis alone. In parallel, one must build transparent legal liability structures that will allow for a quick and fair response in the event of irregularities. Only then will AI be able to become a real support for cybersecurity and not a source of new legal and organisational threats.

5 Towards Responsible AI in Cybersecurity

If artificial intelligence is to permanently reside in cybersecurity systems, it is necessary to develop standards that ensure its responsible and safe use. Although AI offers enormous possibilities—from faster attack detection to automating threat responses—without clear rules for its development and application, it can become a source of new risks instead of protection [23].

One of the key directions of action should be to increase the transparency of AI-based systems. Models that make decisions regarding network security must be as understandable as possible not only to their creators but also to administrators and users. In this context, the concept of so-called explainable artificial intelligence (Explainable AI – XAI) is gaining increasing importance. The goal is to design algorithms in such a way that it is possible to explain why a particular decision was made: why the system marked a specific network traffic as dangerous or why it allowed an unusual connection. The requirement for transparency does not demand sacrificing effectiveness because systems that are well understood enable faster detection of errors and more effective correction methods [14].

The second essential principle for responsible AI implementation in cybersecurity involves continuous monitoring and auditing of algorithm performance. Even the best-designed models can degrade over time, for example, due to changing user behaviours or new types of attacks. Therefore, artificial intelligence-based systems should be regularly monitored for effectiveness, the emergence of errors, or undesirable biases. In practice, this means the necessity of implementing processes for continuous validation and testing of AI performance in a changing environment [17].

The quality of data used for model training stands as an essential factor. The data needs to be representative and diverse while maintaining the lowest possible level of bias. The systems will start to maintain unfair patterns or misjudge threats if the data does not meet these requirements. In the process of building models, it is therefore essential to pay particular attention to where the data comes from, how it is verified, and whether it reflects the reality in which the system is to operate [4].

Finally, it is worth emphasising the importance of education and raising awareness among those who are responsible for implementing and managing AI systems. The treatment of artificial intelligence as an unchallengeable “magic black box” should be avoided. The algorithms’ operational mechanisms together with their boundaries and failure points should be accessible to admins, managers and users. The responsible implementation of technology begins with being aware of its usage [13].

Implementing these principles is not an easy task. It requires the cooperation of many environments: engineers, legal specialists, ethics experts, and system users. But only such an approach gives a chance for artificial intelligence to become a truly effective and safe support for cybersecurity – not a threat, but an ally [8].

6 Case Studies

To better understand the practical challenges associated with the use of artificial intelligence in cybersecurity, it is worth examining specific examples. In this section of the article, two cases will be discussed: one demonstrating the positive application of AI, and the other revealing ethical and legal issues that may arise.

The first example concerns the application of artificial intelligence in protecting a large corporate network from advanced phishing attacks. In one of the European financial sector companies, an AI system based on user behavioral analysis was implemented. The model learned how employees normally communicate via email and how they use the company’s resources. Thanks to this, when cybercriminals attempted to carry out an attack by impersonating superiors and demanding transfers, the system automatically detected unusual behavior patterns and blocked the dangerous messages before they even reached the inboxes. This case showed the enormous benefits that the application of artificial intelligence can bring in situations where humans are unable to independently analyze such large volumes of data in real time. The system required proper implementation to function effectively because the algorithms received continuous monitoring and updates and users received training about AI warning interpretation.

The second case demonstrates how AI systems fail quickly when there is no proper control and inadequate data quality. In one of the public institutions in the United States, an artificial intelligence system was implemented to automatically monitor network traffic and detect threats. Unfortunately, the system was trained on a narrow dataset that did

not account for the diversity of typical user activities in this institution. The result was numerous false alarms – completely legal and harmless activities were incorrectly classified as attacks. The administrators stopped paying attention to warnings because of repeated false alarms which resulted in missing an actual ransomware attack. The institution faced major financial and reputational damage as a result. The case demonstrates why AI implementation requires proper data quality maintenance and model updates for changing conditions and user tools for understanding AI decisions.

The examples show how artificial intelligence serves as an effective cybersecurity protection system but creates new security threats when deployed without proper deployment and regulatory measures. AI deployment requires full responsibility in its implementation which balances technical requirements with ethical and organizational factors to minimize resulting risks.

7 A Proposed Framework for Ethical and Legal AI Deployment

Creating responsible artificial intelligence systems for cybersecurity demands more than excellent technology alone. We need to think broader—about people, processes, and what will happen when something goes wrong. The speed and efficiency of AI operations do not free us from our responsibility to guarantee that its choices remain transparent and fair.

The first step should therefore be to ensure transparency. It's not just about experts understanding how the algorithm works. We need to build systems in which administrators, and to some extent even users, will be able to understand why a particular decision was made. If we do not understand what AI is doing, it will be difficult to trust it in crisis situations.

The second thing is responsibility. When implementing artificial intelligence, it is necessary to clearly indicate who is responsible for its decisions. Is the model creator, the software provider, or the company using the solution responsible? Without such an assignment of responsibility, it will be difficult to resolve disputes, and even more difficult to repair damages.

Furthermore—we must take the data on which models are trained seriously. The accuracy and fairness of AI decisions will suffer when the training data contains biases or incomplete or outdated information. The quality of data along with their diversity and timeliness needs to be monitored at the same level as model effectiveness.

One must not forget that AI systems require constant monitoring. Models that work well today may struggle with new types of threats in a few months. AI needs to be regularly updated, tested, and improved, otherwise it will become more of a problem than a help.

In summary, the responsible implementation of artificial intelligence is not only a technical challenge but, above all, an organizational and social one. The point is for technologies to truly serve people, rather than creating new risks that we will then be unable to control.

8 Conclusions

The development of artificial intelligence in cybersecurity is a process that simultaneously fascinates and worries. On one hand, we are dealing with tools that genuinely enhance our ability to defend against increasingly advanced threats. On the other hand, we observe how quickly new problems arise for which we do not yet have ready answers.

AI delivers substantial value to security systems despite its undeniable benefits. The system enables quick attack detection and real-time response capabilities while processing unmanageable amounts of data that exceeds human capabilities. Each achievement from these successes requires a corresponding expense. The serious risks of model-based automatic decisions and operation transparency issues and accountability challenges must be addressed without delay.

The article demonstrates that technology effectiveness should not disregard the ethical and legal outcomes of its implementation. AI systems used in cybersecurity need to achieve efficiency while maintaining fairness and transparency and upholding user protection principles. Building responsible systems demands a comprehensive approach which unites technological development with legal accountability and social impact assessment.

Looking at how rapidly artificial intelligence is developing, one thing is certain: the sooner we set clear boundaries and standards of responsibility, the better prepared we will be for the challenges that lie ahead. Ultimately, technology is just a tool – it is up to people whether it becomes a source of security or another crisis of trust in the digital world.

References

1. Adu-Kyere, A., Nigussie, E., Isoaho, J.: Analyzing the effectiveness of IDS/IPS in real-time with a custom in-vehicle design. *Procedia Comput. Sci.* **238**, 175–183 (2024)
2. Alsafwani, N., Fazea, Y., Alnajjar, F.: Strategic approaches in network communication and information security risk assessment. *Information* **15**(6), 353 (2024)
3. Awadallah, A., et al.: Artificial intelligence-based cybersecurity for the metaverse: research challenges and opportunities. *IEEE Commun. Surv. Tutorials* (2024)
4. Bran, E., Rughiniş, R., Turcanu, D., Radovici, A.: AI leads, cybersecurity follows: unveiling research priorities in sustainable development goal-relevant technologies across nations. *Sustainability* **16**(20), 8886 (2024)
5. Dey, S.M., Dutta, P.K.: From code to care and navigating ethical challenges in AI healthcare. In: *Human-Centered Approaches in Industry 5.0: Human-Machine Interaction, Virtual Reality Training, and Customer Sentiment Analysis*, pp. 210–225. IGI Global Scientific Publishing (2024)
6. Dube, R.: Faulty use of the cic-ids 2017 dataset in information security research. *J. Comput. Virol. Hacking Tech.* **20**(1), 203–211 (2024)
7. Galinec, D., Steingartner, W., Combining cybersecurity and cyber defense to achieve cyber resilience. In: *2017 IEEE 14th International Scientific Conference on Informatics*, p. 87–93. IEEE (2017)
8. Gauthier-Maxence, C.: European cybersecurity and AI framework: towards proactive regulation for a secure digital future. *EU Law Live* **212**, 1–8 (2024)

9. González, A.L., Moreno, M., Román, A.C.M., Fernández, Y.H., Pérez, N.C.: Ethics in artificial intelligence: an approach to cybersecurity. *Intel. Artif.* **27**(73), 38–54 (2024)
10. Hashmi, E., Yamin, M.M., Yayilgan, S.Y.: Securing tomorrow: a comprehensive survey on the synergy of Artificial Intelligence and information security. *AI Ethics* **5**(3), 1911–1929 (2024)
11. He, Z., Homayoun, H., Sayadi, H.: Guarding against the unknown: deep transfer learning for hardware image-based malware detection. *J. Hardware Syst. Secur.* **8**(2), 61–78 (2024)
12. Hoenig, A., Roy, K., Acquaah, Y., Yi, S., Desai, S.: Explainable AI for cyber-physical systems: issues and challenges. *IEEE Access* **12**, 73113–73140 (2024)
13. Humphreys, D., Koay, A., Desmond, D., Mealy, E.: AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business. *AI Ethics* **4**(3), 791–804 (2024)
14. Ilieva, R., Stoilova, G.: Challenges of AI-driven cybersecurity. In: 2024 XXXIII International Scientific Conference Electronics (ET), pp. 1–4. IEEE (2024)
15. Islam, M.T., Syfullah, M.K., Rashed, M.G., Das, D.: Bridging the gap: advancing the transparency and trustworthiness of network intrusion detection with explainable AI. *Int. J. Mach. Learn. Cybern.* **15**(11), 5337–5360 (2024)
16. Kasula, B.Y.: Ethical implications and future prospects of artificial intelligence in healthcare: a research synthesis. *Int. Meridian J.* **6**(6), 1–7 (2024)
17. Kaushik, K., Khan, A., Kumari, A., Sharma, I., Dubey, R.: Ethical considerations in AI-based cybersecurity. In: Next-generation cybersecurity: AI, ML, and Blockchain, pp. 437–470. Springer, Singapore (2024)
18. Kubanek, M., Szymoniak, S.: Ethical challenges in AI integration: a comprehensive review of bias, privacy, and accountability issues. *Lead. Role Smart Ethics Digital World*, 75–85 (2024)
19. Kumar, D., Suthar, N.: Ethical and legal challenges of AI in marketing: an exploration of solutions. *J. Inf. Commun. Ethics Soc.* **22**(1), 124–144 (2024)
20. Kun, E.: Searching for the appropriate legal basis for personal data processing for cybersecurity purposes under the NIS 2 Directive: legal obligation and/or legitimate interest? *Comput. Law Secur. Rev.* **56**, 106098 (2025)
21. Le Merrer, E., Pons, R., Trédan, G.: Algorithmic audits of algorithms, and the law. *AI Ethics* **4**(4), 1365–1375 (2024)
22. Machado, H., Silva, S., Neiva, L.: Publics' views on ethical challenges of artificial intelligence: a scoping review. *AI Ethics* **5**(1), 139–167 (2025)
23. Malatji, M., Tolah, A.: Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. *AI Ethics* **5**(2), 883–910 (2024)
24. Mienye, I.D., Swart, T.G.: ChatGPT in education: a review of ethical challenges and approaches to enhancing transparency and privacy. *Procedia Comput. Sci.* **254**, 181–190 (2025)
25. Perumal, A.P., Chintale, P., Molleti, R., Desaboyina, G.: Risk assessment of artificial intelligence systems in cybersecurity. *Am. J. Sci. Learn. Dev.* **3**(7), 49–60 (2024)
26. Upadhayay, Y., Sharma, R.: Cybersecurity and legal considerations in AI applications for national security. *Educ. Adm. Theor. Pract.* **30**(5), 11797–11803 (2024)
27. Sarker, I.H., Janicke, H., Mohsin, A., Gill, A., Maglaras, L.: Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects. *ICT Express* **10**(4), 935–958 (2024)
28. Shaukat, K., Luo, S., Varadharajan, V.: A novel machine learning approach for detecting first-time-appeared malware. *Eng. Appl. Artif. Intell.* **131**, 107801 (2024)
29. Szymoniak, S., Kubanek, M.: Biometry-based verification system with symmetric key generation method for internet of things environments. *Sci. Rep.* **15**(1), 5464 (2025)
30. Szymoniak, S., Siedlecka-Lamch, O.: Securing meetings in D2D IoT systems. *ETHICOMP* **2022**, 31 (2022)

31. Szymoniak, S.: Using a security protocol to protect against false links. In: *Moving Technology Ethics at the Forefront of Society, Organisations and Governments*. Universidad de La Rioja, pp. 513–525 (2021)
32. Vashishth, T.K., Sharma, V., Samania, B., Sharma, R., Singh, S., Jajoria, P.: Ethical and legal implications of AI in cybersecurity. In: *Machine Intelligence Applications in Cyber-Risk Management*, pp. 387–414. IGI Global Scientific Publishing (2025)
33. Vulpe, S.N., Rughiniş, R., Țurcanu, D., Rosner, D.: AI and cybersecurity: a risk society perspective. *Front. Comput. Sci.* **6**, 1462250 (2024)
34. Yuan, S., Yang, M., Reniers, G.: Integrated process safety and process security risk assessment of industrial cyber-physical systems in chemical plants. *Comput. Ind.* **155**, 104056 (2024)
35. Zuwanda, Z.S., Lubis, A.F., Solapari, N., Sakmaf, M.S., Triyantoro, A.: Ethical and legal analysis of artificial intelligence systems in law enforcement with a study of potential human rights violations in Indonesia. *East J. Law Hum. Rights* **2**(03), 176–185 (2024)



Cybersecurity, Best Practices: A Comprehensive Guide

Pedro Brandao^(✉) 

Instituto Superior de Tecnologias Avançadas, Lisbon, Portugal
pedro.brandao@istec.pt

Abstract. Due to ever-evolving cyber threats, organizations and individuals are compelled to adopt effective cybersecurity measures to protect sensitive data and digital assets. This comprehensive guide will examine some of the best practices in cybersecurity, including risk assessment, strong authentication methods, data encryption, and network security strategies. Other topics discussed in this paper include employee training, incident response planning, and compliance with cybersecurity frameworks. If implemented, these best practices will significantly reduce the likelihood of businesses and individuals becoming victims of cyberattacks while enhancing overall digital security.

Keywords: Cybersecurity · Incident · Artificial Intelligence

1 Introduction

Welcome to the Cybersecurity Best Practices: A Comprehensive Guide. The ultimate goal of this document is to provide a holistic view of best practices and concepts relevant to information security while addressing the needs of small businesses and technical teams. Derived from the insights of cybersecurity experts, both general guidance and technical configuration steps are provided to implement industry standards and best practices effectively.

There is no exaggeration in claiming the importance of security. Systems fail individuals, schools, corporations, and government departments every day. Not only is it crucial, but it is also profitable. There are no indicators suggesting that the number of data-related breaches is declining; in fact, all indications suggest the opposite. Proof of this comes from the data breaches that have already exposed individual records. It falls to cybersecurity professionals to prevent damage and to assist where normal practices fail.

The richness of the internet and the digital world in technology is limitless. Every day, new developments in technology assist in enhancing inventions and the creation of valuable digital systems or devices. Attacks and threats are driven by motives that vary widely. Knowledge is strength; therefore, the more knowledge professionals have about the motives driving adversaries to launch an attack, the better their defense strategy can be.

2 Understanding Cybersecurity

Throughout human history, there have been countless innovative inventions and technologies. Each has brought about significant social change and numerous benefits to society, improving lives in many ways. Unfortunately, humanity has also gained the ability to inflict great harm on itself using technological advancements, as demonstrated in WWII with the advent of nuclear weapons. As innovation continues to accelerate globally, security standards must rise alongside it to protect valuable digital data. Just as governments around the world strive to prevent acts of terror, there is also a growing need to safeguard infrastructure against cyber attacks [1]. This is especially true in the IT industry, where corporations are entrusted with personal, financial, and proprietary data. Understanding cybersecurity is essential for protecting such information from today's most common cyber threats.

Information security measures are designed to protect the confidentiality, integrity, and availability of data. An organization's information security asset consists of a combination of technical and organizational measures, routines, and best practices that are employed to manage risks pertinent to the organization. A security asset can operate at the collective, asset-specific, or modal layer. A collective security asset encompasses, for instance, all the security measures of a single company. In contrast, an asset-specific one may include related items such as equipment and software in a network environment. At the modal layer, a security asset refers to the means that enable access to or utilization of the asset-specific measures. This can include, for example, training, installation manuals, and access rights. Training serves as both a preventive measure, since competent use can avert many risks, and as a means of access, as having comprehensive training materials may be a fundamental requirement for optimal asset-specific security behavior.

From an organizational perspective, the modal layer can be quite broad, potentially encompassing the entire field of education and training. However, this discussion examines the issues from the viewpoint of an information security professional responsible for safeguarding the information security of a specific collective asset, such as a single company, authority, or other organization. Properly implementing modal-level security can be very powerful if it is set up and executed correctly. Therefore, managing it is a crucial part of securing the assets. A well-structured modal asset can significantly mitigate problems arising from neglecting an asset-specific concern.

3 Cyber Threat Landscape

Rapid developments in cyberspace in recent years have introduced significant risks to society, including national interests. The cyber threat landscape has become even more complex in the gray zone, posing a considerable challenge to national security governance capabilities worldwide. Most in-depth cyber-attacks are covert and have noticeable effects that fall short of outright measured engagement, yet they are still clearly intended to undermine an adversary's interests [2]. Business organizations facing the dual threats of network security and intellectual property security also encounter increased risks to their survival and development. Rapid developments in cyberspace in recent years have introduced significant risks to society, including national interests.

The two zones typically refer to the areas located between peace and war and encompass three specific periods.

The most recent development is recognized as fifth-generation warfare, extending beyond traditional combat. The cyber threat landscape is increasingly complex in the gray zone and presents a significant challenge to governance capacities regarding national security worldwide. This ultimately impacts the relevance of current network security practices. A high degree of connectivity among states and international organizations directly and indirectly faces network security threats, resulting in serious consequences for economic growth and national security. Scenarios facilitate risk assessment, the provision of well-founded policy advice, and the evaluation of existing or the development of new policy options and resources.

Cyber threats come in various forms, all designed to gain unauthorized access to systems and data, or to steal or otherwise carry out malicious actions against individuals or information [3]. Different types of cyber threats exist. Hacktivism is a type of cyber threat where the attacker has a political or social agenda that seeks to disrupt the target's operations or spread harmful information about the target organization or its customers. Phishing is an Internet-based cyber threat that involves deceiving potential victims into clicking on a seemingly legitimate link or email that actually directs them to an infected website, file, or email attachment containing malicious code meant to compromise the user's computer. Malware is a category of cyber threat that aims to infect a computer system to gain unauthorized access to files, install additional threats, or execute automated actions such as copying emails, passwords, or keystrokes. Malware appears in many forms, including viruses, worms, trojans, ransomware, adware, backdoors, and more. Additionally, certain cyber threats like DDoS attacks can easily render a system's network, website, or cloud service non-operational simply by overwhelming it with requests. Security risks from people include, but are not limited to, industrial espionage, recruiting spies or rogue employees, social engineering, and conspiracy spreading from disgruntled employees.

4 Security Policies and Procedures

Security policies are a set of rules and procedures designed to enhance an organization's security. Strong security policies help identify and block potential threats, ensuring that confidential information is not inappropriately distributed or viewed. It is crucial to regularly update and audit security policies [7]. Policies and procedures are developed to protect not just paper or computer-based information, but also information stored in other forms, such as conversations, physical security of information, and personal access to or viewing of information. Disciplinary actions for breaches of security policy are generally at the employer's discretion. A code of good practice for information security consists of mandatory guidelines for information security. Health and safety legal requirements complement the policy's provisions. Computer log-in information is always confidential and should not be shared; it must not be written down in a recognizable form. Ultimately, employees are responsible for maintaining the confidentiality of log-in information [4]. Unnecessary printouts of reports containing personal information should always be shredded rather than discarded with regular waste. To protect

network security, unauthorized software installation should never occur. An audit trail of all transactions affected by each password used is automatically generated for future review.

It is always helpful to appoint an information technology security officer to develop, monitor, audit, and amend security measures. The officer should also instill in employees the importance of having an appropriate level of awareness regarding the sensitivity of information assets. A good policy principle is to restrict access where employees do not need it. Access to data processors is limited to relevant parts of the information system pertaining to their responsibilities. Access is granted only to the data or tables necessary for job functions. Additionally, a clear audit trail is maintained to track what data has been accessed. Finally, network traffic is monitored for performance reasons. All personal information is encrypted before transmission. Processes are in place to ensure that data on stolen laptops and other devices cannot be accessed. Direct marketing permission does not extend to data sharing. The email system complies with legal requirements to identify contact data in each message. Telephone preference service files are checked monthly. Faxed information is not accepted from external sources.

Developing Security Policies ~ Develop Information Security Policy ~ [7]: The first step in developing an information security policy is to secure management support and guidance in the policy's creation. Management should view policy development as a strategic task that safeguards organizational information assets. The next step involves assessing the existing policies and procedures, which is essential as it allows the organization to identify gaps in current policies and determine whether they effectively address risks. Assessing existing policies and procedures ensures that new policies comply with established standards, thus increasing the likelihood of successful implementation within the organization. The assessment process helps gather critical materials, such as existing policy and procedure documents. Therefore, assessment should be the starting point in developing an information security policy. Creating such a policy needs to consider standards relevant to the size and type of services the organization offers. The goal of this best practice is to equip start-up SMEs with the information necessary to select standards for an information security management system (ISMS) and implement them. The information within this best practice aims to focus on international standards exclusively. International standards are widely accepted and recognized in the global business community. Choosing standards is a vital decision in developing a security policy, as they dictate the nature and quality of the security measures. A decision must first be made to engage with a specific standard. Next, the relevant standard should be purchased, and a team comprised of individuals with a strong grasp of information security should be assembled.

The Conclusions of an Organization Systems Security Policy ~ [4]: This Organization Systems Security Policy is a written document created by the organization's management that outlines its stance on the security of information and IT systems. It specifies protected services and data while identifying threats to these systems. The policy serves as a minimum standard for experts and management to establish and manage the Information Security Policy. Certain critical systems in the new Digital Workspace environment outline the requirements for all connected devices and data exchanges, as specified by the security policy. Data is classified into three levels: Open level information includes

public or copyrighted material, unregistered trade secrets, or company confidential data; official registered trade secrets and authorized sensitive topics; and secret instructions or highly sensitive organizational information. Distributed information, in any format, must be protected by transmission encryption. Regular data backups are essential for restoring lost or compromised data. The FSM IT can provide backup services, or an appropriate internal routine can be established. Unintentional data loss can occur due to technical problems, crashes, or the loss or theft of devices. Backup and recovery procedures must be documented and cannot rely solely on IT Services; it is necessary to have a plan in place to restore lost data within a specified timeframe. General measures should be implemented, including basic rules such as shutting down the machine ASAP. All security incidents that could compromise system security must be documented, and the responsible party may face action if they violate the policy. Monitoring the security of connected devices is crucial, and proper infrastructure ensures real-time, automatic processes and accurate, periodic collection of usage logs. Monitoring is also valuable for ensuring compliance. Changes to system configurations are nearly impossible without leaving traces. Unauthorized intrusions can be swiftly detected, and defense mechanisms can be implemented to prevent further actions. An external expert organization has been assigned to conduct an annual examination of these systems, performing penetration and vulnerability tests on connected devices.

5 Data Protection

Data protection starts with identifying the most critical data, determining who has access, outlining how the data will be protected, and establishing methods for detecting potential data breaches. Once the critical data is secured through encryption, monitoring tools can analyze network traffic, allowing for prompt detection of suspicious activity. Patches should be applied quickly to software with security vulnerabilities, and using virtual private networks or dedicated communication lines is preferable when transferring data between locations. Additionally, it is crucial to ensure that the supplier's cybersecurity procedures continually meet established requirements. Common issues for newcomers to IT security include overlooking cybersecurity when introducing new equipment or services, failing to provide IT equipment and services robust enough to meet information security standards, and using equipment bought from the retail market without implementing technical protections. A primary goal is ensuring the availability of secure IT services as needed. Significant challenges to achieving this goal include managing the implementation of an information security management system, ensuring the physical protection of data-center hosts, and defining secure recovery strategies to minimize the impact of security incidents on IT service delivery. It is important to note that the secure delivery of IT services is closely linked to applying information security criteria to all IT service acquisitions or internal developments and ensuring ongoing conformity.

The basic concepts of information security are simple yet important. They take the form of three questions: What will be done to maintain the confidentiality of a resource (for example, a server, file, or document)? What is required to maintain its integrity? And how will its availability be ensured? These three questions form the foundation of an approach known as the company's information security policy, which emphasizes

confidentiality, integrity, and availability (CIA). At this stage, the first step is to critically assess the considerations selected for protection, whether they involve a document or an IT service. Here, the starting point is a list of all assets that require protection. Assets can pertain to data or IT services. A risk evaluation method must be used to identify the primary risks to these assets. The result will be a list specifying the appropriate assets and the main risks affecting them. Clarifying the ownership of the identified risks is also essential. Be realistic when assigning ownership within the company, and if necessary, consult stakeholders when determining who should own a risk. It is crucial to confirm the list of risks and their owners, as these should be present across all cases or IT systems.

Encryption is the process of transforming plaintext data into a non-readable form, known as “ciphertext.” When sensitive data is encrypted, it is altered in such a way that without the encryption key, it remains undecipherable. The encryption key is essential to revert the data back to its readable (plaintext) form. At System and Organization Controls (SOC)1, a ubiquitous computing environment is assumed where multiple entities may access the same databases. These entities may have varying data access privileges, and there is typically a need to restrict specific data elements for certain roles. An automatic data masking scheme must be integrated into applications within this environment. Production databases should remain in an uncorrupted state after the masking process occurs. It is necessary to maintain a link between the original data elements and the masked data for traceability [10]. Conversely, nonproduction databases are transient and need to be shielded from sensitive data. Data masking at the output level may encounter data format issues and data aggregation problems. A new metric is proposed to measure the information lost during masking. In a production environment, input-based data masking schemes become barriers that prevent continuous monitoring of data access. Masking algorithms must have an implementation time significantly shorter than the offline runtime on any database table.

6 Incident Response

Upon detecting a security incident, it is critical to assign a primary incident responder. An event may require establishing a communication system to report the scope, location, and duration of the incident, the types of systems involved, the data sampled, and the technical environment where the data resides. The need for a communication system may vary depending on the incident type and the sensitivity of the data. This communication system must accommodate the need for multiple communication mediums [11]. It may be necessary to deploy incident responders to outreach locations or remote sites. Control must be maintained regarding who is sent where, with what resources, and their mission assignments. The purpose of an incident response plan is to provide guidance for detecting, responding to, minimizing, and resolving acute security incidents. The Incident Commander (IC) is the individual responsible for managing the incident from start to finish. Several rank-and-file personnel involved in a computer security incident may be unfamiliar with the principles or practices of an incident response team. The Public Information Officer will only be utilized in breaches involving outside agencies or the public. In smaller events, this may be folded into the IC’s responsibilities. The Liaison Officer serves as a central point of contact and manages requests between the responding

organization and those seeking assistance. This position is typically not utilized but may be necessary in cases requiring law enforcement involvement or coordination between partner organizations. Operations manage the tactical operations of the incident response (IR) organization and are responsible for conducting the actual response activities.

Most small to medium-sized organizations lack the ability to effectively address incidents involving cyber-attacks. Those that do possess this ability are often poorly trained and inadequately equipped. Having a plan designed to mitigate potential breaches can significantly reduce possible damages. These damages may include fines, loss of reputation, loss of income, and civil suits. Many states have enacted laws regarding breaches of confidentiality relating to patient medical information. Additionally, numerous laws also pertain to financial information. Financial institutions and merchants that process credit card data are likewise subject to regulations governing breaches of financial information. An acceptable and current strategy should be developed by a Cyber Incident Response Team (CIRT) or a consultant specialized in creating acceptable plans. Currently, incidents are managed on an ad hoc basis. Typically, a member of the Information Technology staff discovers a situation after it has occurred. This staff member then investigates the issue, often incorrectly. Such errors can compromise potential forensic data that may reveal hacking attempts and those responsible. This document will discuss the need for a properly trained, staffed, equipped, and documented plan for a CIRT. The goal is to respond swiftly to an incident or breach, minimizing losses while providing forensic data for notifications, disciplinary actions, legal proceedings, risk mitigation, or identification of zero-day exploits. Properly trained emergency service personnel, utilizing the Incident Command System model, could establish a partially effective CIRT overnight capable of addressing multiple concurrent events and scaling to any size disaster.

7 Emerging Technologies in Cybersecurity

This is a comprehensive guide to best practices in cybersecurity. It covers twelve critical areas of knowledge and insight for both professionals and casual observers in the field of cybersecurity. Domains are broad categories that encompass everything from availability and access to disaster recovery and organizational operations. Practices are specific methods used to achieve agreed-upon objectives in the realm of security. Although practices vary by organization and circumstance, there are universal considerations to keep in mind throughout this reading. [13]. Methods are mechanisms and procedures used to accomplish security tasks, including common protocols, data types, endpoints, and gateways. Emerging technologies refer to new, compelling, and untested security approaches, products, programs, or vendors that have recently gained attention or funding. These technologies can be highly specific or incredibly broad, and they are vital components of the rapidly changing digital landscape. Security methods often lag behind widely adopted technologies, and emerging technologies in the security field can provide organizations with essential means to thwart threats before they become widely exploited vulnerabilities. Below, four emerging technology providers will be analyzed closely, in addition to a broad discussion of such options.

Powerful AI and machine learning tools can help secure your network and digital infrastructure. This guide describes them in plain English, without using mysterious

or technical terms, and suggests readiness steps for quickly integrating them into your cybersecurity strategy: - IT teams often struggle to identify all potential cyber threats on their networks effectively and quickly, which explains why attackers typically remain undetected for an average of 205 days. An attacker only needs hours or days to exploit this physiological limitation against their targets. Equipped with powerful AI, attackers have crafted even more Advanced Persistent Threats that have plagued businesses, infrastructure, and economies with stealthy intrusions never seen before. - Red Teams will implement AI-assisted cyberattacks using datasets that cover various aspects of their targets. It is essential to limit access to critical digital infrastructure, such as IoT control systems, which are still easy to find and access. Controllers, which dictate actions in cyber-physical systems, gather data stored and transmitted from sensors and employed controls. - Blue Teams utilize datasets that provide detailed insights into the operations of digital industries and smart infrastructures. A comprehensive overview is offered on actual datasets that are realistic to envision and utilize in defense, including some that have not yet been exploited. Readiness steps are recommended to quickly deploy AI-enhanced defenses and keep pace with sophisticated AI-assisted attacks. Actionable examples are provided to make the advice practical.

8 Ethics and Artificial Intelligence

In recent years, artificial intelligence has shifted from a futuristic concept to an integral part of our daily lives. We encounter it in the recommendations we receive online, the responses from our smartphones, and even in decisions made by governments and large corporations. As this technology becomes more autonomous and powerful, an important question arises: how can we ensure that AI acts ethically, respecting human values and social justice?

Ethics in AI is not just a theoretical exercise; it is a pressing concern that affects the very structure of our societies. Developers and policymakers are now being urged to reflect deeply on how these systems are created and deployed. It is insufficient for AI to be efficient—it must also be fair, transparent, and accountable.

One of the first ethical principles that comes to mind is transparency. People need to understand how an AI system makes its decisions, especially when those decisions impact lives—as in judicial rulings or medical diagnoses. Without this clarity, trust erodes, and AI becomes a “black box”—a mysterious mechanism whose inner workings remain opaque.

Closely related to transparency is the principle of accountability. Who is responsible if an AI system discriminates against someone during a hiring process, or if a self-driving car causes an accident? Assigning responsibility is crucial to ensure that technological development stays within legal and moral boundaries.

Fairness also plays a central role. AI systems learn from data, and if that data reflects social inequalities, these inequalities can be perpetuated or even amplified. Documented cases exist where AI tools used in judicial systems or hiring processes have shown racial or gender biases—not because the programmers intended it, but because the data they utilized was biased from the start.

Then there's the question of privacy. In a world where AI thrives on vast amounts of personal data, the lines between public and private are becoming increasingly blurred.

Surveillance technologies, especially facial recognition, have ignited fierce debates worldwide. In some areas, their use in public spaces has already been banned or severely restricted.

The autonomy of AI systems raises fundamental questions. As machines take over tasks once reserved for humans, concerns about job displacement increase. What roles will remain for people in an automated society? Will AI complement human labor or replace it entirely?

Furthermore, the weaponization of AI introduces a darker aspect to the debate. Autonomous drones, cyberweapons, and AI-driven surveillance can be employed not only for protection but also for control and harm, raising significant ethical and geopolitical concerns.

To address these challenges, various organizations have proposed ethical frameworks. The European Union, for instance, published its guidelines for trustworthy AI, emphasizing human-centric values such as robustness, transparency, and diversity. The OECD and IEEE have also promoted international principles aimed at ensuring that AI supports human well-being and development.

Real-world examples offer sobering insights. Tools like the COMPAS algorithm in the United States, designed to predict the likelihood of a defendant reoffending, were found to disproportionately classify Black defendants as high risk. Meanwhile, Amazon had to abandon its AI recruiting tool after discovering that it systematically favored male candidates over equally qualified women.

These cases illustrate why ethics cannot be an afterthought. They must be integrated into the design process from the very beginning. This necessitates interdisciplinary collaboration. Engineers, ethicists, psychologists, and lawyers must unite to shape AI that is not only intelligent but also wise.

Ultimately, the aim is not to halt AI, but to steer it—ensuring that its power is utilized responsibly, transparently, and humanely. The future of AI should be founded not just on lines of code but on the values we all share.

9 Concluding Remarks

Overall, it has become clear that information security is not just about technology. It is a vast field that strongly interfaces with business processes, legal practices, and individual behaviors [4]. While information security best practices can seem elusive, especially when trying to connect them into a coherent entity, this one-page cheat sheet aims to simplify things. It seeks to avoid lengthy essays that discuss why certain aspects are important and instead summarizes actionable steps. This approach enables a shift in thinking about information security from a business risk perspective. Naturally, one could explore each topic in greater depth, but keep in mind that the primary goal is to teach you how to fish, rather than to provide a ready meal.

References

1. Borky, J.M., Bradley, T.H.: Protecting Information with Cybersecurity (2018). ncbi.nlm.nih.gov

2. Altaher Ben Naseir, M., Dogan, H., Tersoo Apeh, E., Richardson, C., et al.: Contextualising the National Cyber Security Capacity in an Unstable Environment: A Spring Land Case Study (2019)
3. Schmitt, M.: Mobile Security for the modern CEO: Attacks, Mitigations, and Future Trends (2022)
4. Kaila, U., Nyman, L.: Information Security Best Practices: First Steps for Startups and SMEs (2018)
5. Murtaza, M.B.: Developing An IT Risk Assessment Framework (2007)
6. Nurse, J.R.C., Creese, S., De Roure, D.: Security risk assessment in Internet of Things systems (2017)
7. Alshaikh, M., Maynard, S.B., Ahmad, A., Chang, S.: Information Security Policy: A Management Practice Perspective (2016)
8. Weiss, M.S.: Network Defense: The Attacks of Today and How Can We Improve? (2009)
9. Assessing the Value of Network Security Technologies. University of Texas at Dallas Cavusoglu
10. Ajayi, O.O., Adebisi, T.O.: Application of Data Masking in Achieving Information Privacy (2014)
11. DeVoe, C., Rahman, S.: Incident Response Plan for a Small to Medium Sized Hospital (2015)
12. Sirur, S., Nurse, J.R.C., Webb, H.: Are we there yet? Understanding the challenges faced in complying with the General Data Protection Regulation (GDPR) (2018)
13. Kelly, T.M.: Whou27s In and Whou27s Out?: Whatu27s Important in the Cyber World? (2016)
14. Garcia, M., Forscey, D., Blute, T.: Beyond the Network: A Holistic Perspective on State Cybersecurity Governance (2017)



Enhancing Health Information Access via ChatGPT and the E-Citizens Portal

Natalija Zeko¹, Darko Galinec¹ , and William Steingartner²  

¹ Zagreb University of Applied Sciences, Zagreb, Croatia
darko.galinec@tvz.hr

² Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Košice, Slovakia
william.steingartner@tuke.sk

Abstract. With the development of technology and digital systems, everyday life is undergoing significant changes. As one of the most significant technological advances of today, dating back to the mid-20th century, artificial intelligence (AI) is now used in various industries, especially in healthcare. Digitalization enables faster and more efficient access to information, and the advancement of artificial intelligence brings new opportunities for improving healthcare services. This paper deals with the application of artificial intelligence, especially ChatGPT, in the healthcare sector through the e-Doctor web application. The theoretical part provides insight into the basic concepts of artificial intelligence, its history, types and applications with an emphasis on the healthcare sector. The paper presents diagrams to illustrate the structure of the system, which includes business processes, data models and user activities. The practical part of the paper focuses on the development of a prototype of the e-Doctor web application that uses ChatGPT. ChatGPT is a virtual assistant that every user of the system can contact when they have certain questions. The aim of this paper is to demonstrate the integration of ChatGPT with the e-Citizens portal through the development of a prototype of the e-Doctor web application. The e-Doctor web application provides an insight into how artificial intelligence can improve access to health information and facilitate communication. Thus, the application provides functionalities such as reviewing medical data, asking questions to a virtual assistant, and generating personalized diet plans. The main goal of the application is to reduce the administrative burden on healthcare workers and improve the quality of healthcare services. The outcome of the research is a functional prototype of the e-Doctor web application for improving access to health information.

Keywords: artificial intelligence · ChatGPT · healthcare · virtual assistant · web application · implementation · PHP · MySQL

1 Introduction

The term intelligence refers to the cognitive skills and thinking abilities of both humans and animals, also known as natural intelligence. The human brain has

the ability to learn, reason, think, remember, and pay attention, and these abilities are called cognitive skills. Artificial Intelligence as a term consists of two words: “artificial” and “intelligence”. The word “artificial” refers to something that is not actually natural or made by humans. AI can be called “machine intelligence” which is demonstrated by a computer or electronic machine [18, 19]. AI creates systems that imitate human cognitive abilities, which include learning, problem solving, and decision-making, which gives the ability to information processing systems similar to human thinking. It can be said that AI is a field of computer science that aims to simulate cognitive processes on computers [8]. Thus, there is a need for a computer to implement artificial intelligence into its system in order to demonstrate its capabilities [12, 20]. Today, artificial intelligence is present in many areas such as: healthcare, agriculture, retail, industry, manufacturing, and insurance and finance [5, 15]. The application of artificial intelligence is reflected in companies that use it in internal operations to improve business growth, and on the other hand, there are companies that provide new AI solutions. In healthcare, a large application is reflected in which artificial intelligence makes a better and faster diagnosis than medical staff. The idea of AI reshaping the medical profession is also reflected in literature, such as Robin Cook’s novel *Cell* (2014), where a smartphone-based system called iDoc functions as a fully capable personal physician, surpassing human doctors. This fictional scenario mirrors real-world discussions about the evolving role of AI in healthcare, as examined through the e-Doctor application in this paper.

The general objective of this study is to briefly describe AI, and the specific objective is to explore and demonstrate the possibilities of its application in the field of healthcare [4]. So generally, it’s about to develop a functional prototype of the e-Doctor web application by integrating the ChatGPT language model into the governmental e-Citizens platform, with the aim of exploring the potential of artificial intelligence to enhance access to health information, reduce the administrative burden on healthcare professionals, and improve user experience in digital public health services. Specifically, we aim to describe the architecture and technical components of the e-Doctor application, detailing the technologies, processes, and interaction flows involved, to implement and test a functional prototype that enables patient interaction with the virtual assistant VirtualDoc, powered by ChatGPT. In future work, the aim is to integrate the prototype with the e-Citizens portal, addressing the technical and security requirements, with particular attention to enhance data privacy and strong digital authentication as the first barrier of cyberattack regarding in users’ regard.

The structure of the paper is the following: in Sect. 2, the types of artificial intelligence in context of applications in healthcare are discussed. Section 3 covers the description of our prototype of the web application for improving the access to health information. Finally, the Sect. 4 concludes our paper.

2 Types of Artificial Intelligence

The fundamental concept before the advent of computers is precisely artificial intelligence, which is applied in many different aspects of business as well as

domains for process optimization and problem solving. This technology can be divided according to different criteria, some of which are the ability to perform tasks without supervision and the ability to develop cognitive functions, as well as reasoning in a human-like manner. Therefore, artificial intelligence encompasses two basic categories:

1. AI Types Based on capabilities
2. AI Types Based on functionalities.

2.1 Branches of Artificial Intelligence

Artificial intelligence that aims to imitate human behavior is divided into several types (see Fig. 1), the most famous of which is machine learning, which enables systems to learn and improve from experience without explicit programming. Thus, machine learning without explicit programming enables systems to learn and improve their experiences as well as for machines to process, interpret and use data to solve real-world problems [10]. Furthermore, machine learning algorithms are made of mathematical calculations that are encoded in the so-called machine-readable language that enables classification and prediction of outcomes from huge data sets. The characteristics of machine learning are recognizable by examples, and some of them are autonomous vehicles, demand forecasting, image and speech recognition systems, etc. Machine learning can further be divided into three types, namely: supervised learning, unsupervised learning and reinforcement learning [9].

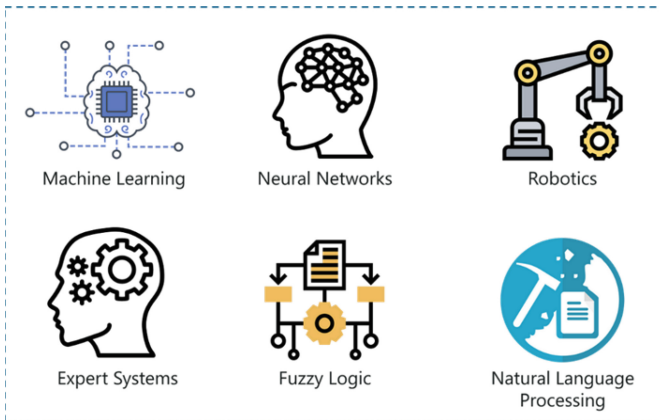


Fig. 1. Branches of Artificial Intelligence (reprinted from [7])

Another branch of artificial intelligence is deep learning, which uses neural networks inspired by the human brain, reducing human intervention and requiring large amounts of data. Amazon Alexa and Google Home are examples of

deep learning. Natural Language Processing allows computers to speak words and written text. It is present in many aspects such as digital assistants, chatbots, and spam detection [21]. Robotics is a third branch that is oriented towards the creation of robots that unites mechanical engineering, electrical engineering, and computer science. It is applied in production lines as well as space expeditions, healthcare, manufacturing, etc. Process of dealing with uncertain and imprecise information where degrees of truth are assessed rather than binary truths or lies. This provides a better way to imitate human thinking, which is a characteristic of fuzzy logic [9, 16].

2.2 Application of Artificial Intelligence in Healthcare

The way AI is used in healthcare has the following features.

1. Artificial Intelligence (AI): Refers to a broad area that includes various techniques and methods for processing and analyzing data.
2. Data: The central part of the diagram shows the different types of medical data used:
 - EMR Data: Data from electronic medical records.
 - Imaging, Genetic, EP Data: Refers to medical images, genetic information, and electrophysiological data.
3. Machine Learning: Technology that allows AI to learn from data without explicit programming. This component takes data as input and generates outputs that assist in clinical activities.
4. Natural Language Processing (NLP): Technology that allows computers to understand and process human language. It is used to process clinical notes written in human language.
5. Clinical Activities: Include screening, diagnosis, and treatment. Machine learning generates outputs that support these activities.

2.3 Basic Features of ChatGPT

On November 30, 2022, ChatGPT (Generative Pre-trained Transformer) was introduced by OpenAI. This is actually a chatbot based on a language model that allows communication according to one's own wishes, format, style, and rapid engineering where it uses previously given queries and answers. There are currently two models that represent a series of pre-trained transformers:

1. GPT-3.5 — free version
 2. GPT-4 — paid version
- GPT-1 appeared in 2018, and GPT-2 in 2019.

Also, versions are available for different versions of platforms such as Android and iOS [13]. ChatGPT uses the Transformer architecture that processes large amounts of text [1]. It is a deep learning technique for understanding and generating text responses in natural language. The GPT-3 is considered the largest language trained model with 175 billion parameters, and has been trained on 8 million documents and about 10 billion words, which allows it to perform tasks as well as generate well-written text. To make everything work well, reinforcement learning based on human feedback was used for training [22].

2.4 AI Applied in the Development of the Proposed System

When introducing AI technology (Fig. 2), it is necessary to check the documentation that achieves the characteristic of transparency, while autonomy implies the rights of patients and individuals to accept or reject recommendations coming from artificial intelligence. The protection of health data is a characteristic of privacy, which is one of the key components that must not be used without the consent of the individual. Equality is an important item because access should be equal for everyone regardless of age, gender and ethnic affiliation, income, etc. During the decision-making process, artificial intelligence tools must be integrated into the workflow of doctors. Ensuring the validity of AI tools and providing benefits to patients are based on evidence derived from clinical trials. Governance refers to the laws and regulations that regulate the use of data and the regulatory system for educating AI tools. Finally, if certain errors and mistakes occur in the application of artificial intelligence, laws and regulations that allocate responsibilities for losses come into force [21].

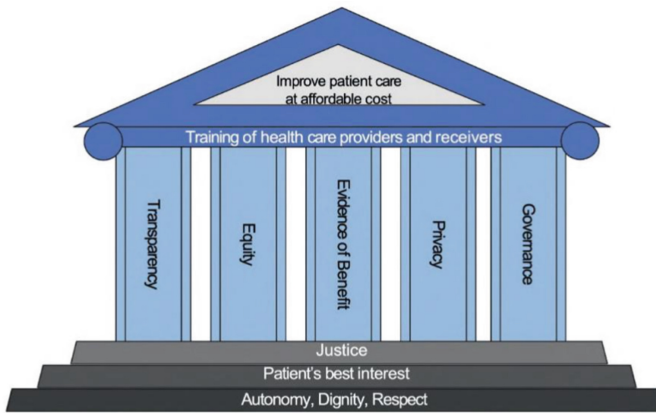


Fig. 2. Implementation of artificial intelligence in healthcare [6]

The application of artificial intelligence in healthcare has resulted in a revolution in medical technology as a computer science that struggles with complex problems, but also many applications in areas with a huge amount of data, and little theory [2]. The term medical technology (English Medical Technology) with the tools it possesses allows patients and society to be provided with the best quality of life with early detection of diseases by healthcare personnel [17]. Classic medical devices have been replaced by new smartphones, wearable devices, sensors and communication systems that contain artificial intelligence. The general population has resulted in a positive attitude towards intelligent medical technologies (English Intelligent Medical Technologies) that enable the so-called 4P model of medicine (English Predictive, Preventive, Personalized and

Participatory) and patient autonomy, and smartphones have become the basis for filling out electronic health records [3,7].

2.5 e-Citizens

e-Citizens is a digital platform developed by the Government of the Republic of Croatia with the aim of providing citizens with simple and efficient access to various public services via the Internet. Access to electronic services offered by various state institutions and bodies is enabled precisely through the e-Citizens system, which was crucial in the modernization of public administration.

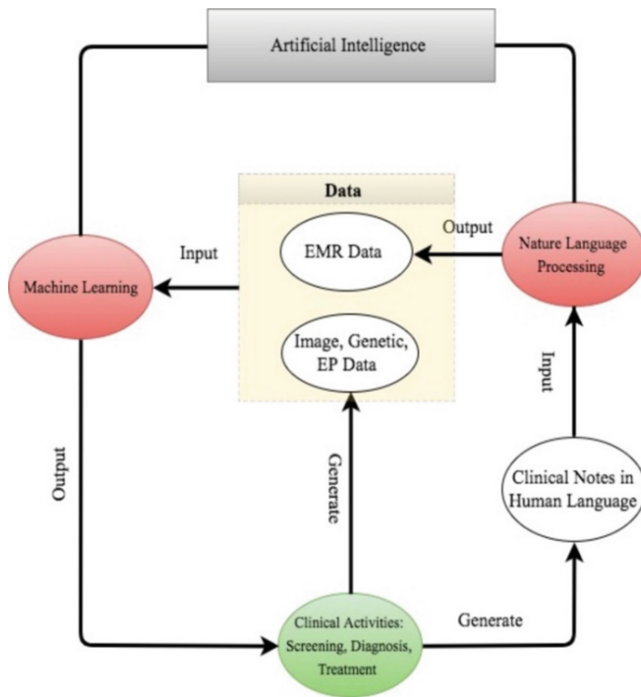


Fig. 3. Depicting the use of artificial intelligence through a flowchart (reprinted from [23])

Key features of the e-Citizens platform are:

- Centralized access to services: quick and easy access to various services without the need to visit physical locations
- Security and privacy: various authentication methods
- Variety of services: e-Matica, e-Tax, e-Health, e-Education, Health Portal, etc.
- Accessibility: the platform is available to all users who can use the service via mobile devices, computers and tablets.

- Ease of use: regardless of digital skills, the platform is easy to use, and has support and guides for users.

3 Prototype Web Application to Improve Access to Health Information

During the creation of the computer model, several programming languages and technologies were used to ensure the functionality, interactivity and adaptability of the application. HTML (HyperText Markup Language) was used for the basic structure and organization of content on the website. CSS (Cascading Style Sheets) is used to stylize elements, that is, to adjust the color, font and appearance of the page. Bootstrap is a front-end framework that offers pre-defined components and styles, which made it possible to quickly create a user interface. A MySQL database was used for data storage and management. In the background part of the application, PHP (PHP Hypertext Preprocessor, Fig. 3) framework Symfony 6 was used, which enabled the overall functionality of the application, such as working with the database, authentication, server-side logic, etc. Also, when implementing chat functionality for VirtualDoc and DocAssist, JavaScript AJAX was used to enable asynchronous sending and receiving of data from the server without the need to reload the page after each question. In this

Table 1. Process List

	Process name	Process description
Authentication and authorization	First login	First login via e-Citizen saves a new user to the database
	User login	Login via the e-Citizen system
	Authorization	Differentiation of access rights between patient and doctor
Medical data review	Medical history	Review of patient's medical history
	Allergies	Review of allergy details and previous reactions
	Findings	Review of lab and test results
	Medications	Review of prescriptions and doses
Interacting with ChatGPT	Chat with ChatGPT	Patients can ask questions related to their health, and doctors can help in diagnosing the patient
Future API integration	Planning Application Programming Interface (API) integration	Planning future data integration with e-Citizen
	Security measures for API integration	Implementation of security measures to protect data during API communication

way, a better and faster user experience is achieved as described in Process List contributed by authors (Table 1).

3.1 Functionality of the e-Doctor Web Application

e-Doctor is a web application developed with the aim of integrating the Chat-GPT AI language model into the e-Citizens system. Namely, the web application is presented as a prototype application, which means that it is in the testing phase. The application is very easy to use with functionalities to contribute to facilitating the healthcare system in the Republic of Croatia. The main goal of the application is precisely to reduce the administrative burden on healthcare staff and speed up the process of providing information to patients. Since it is a prototype, the emphasis is placed on the functionality of the application, while later plans are focused on improving the design. The application uses test data instead of real data from the e-Citizens system and the Health Portal, while in the final version it should retrieve data via API. Also, although the login system currently uses the Symfony authentication system, in the future, login via the National Identification and Authentication System (NIAS) and e-Authorization will be integrated. On the home page, the user can select the “Institutions” option, which displays a list of healthcare institutions with information on the name, address, telephone number and working hours. Each institution allows users to view a list of clinics and their doctors, as well as a schedule of working hours in three shifts (morning, afternoon and night). In the future, it will be possible to view the exact location of the institutions and instructions for getting to them.

3.2 Anonymous User

When an anonymous user arrives at the e-Doctor homepage, they see a modern and simple design that emphasizes quick access to health information. The navigation menu at the top of the page has three options: Home, Institutions, and Login. At the top of the homepage, on the left side of the navigation bar, there is the e-Doctor application logo (Fig. 4). The central part of the page contains a message that emphasizes how the application uses artificial intelligence to make it easier to search for medical information and ask health-related questions. Below the aforementioned text, the user has the option to select the “Login to the system” button, which takes them to the authentication system. The anonymous user can select the “Institutions” option where they see a list of health institutions. The list of institutions lists details about each health center, such as: institution name, address, telephone number, and opening hours. In further implementation, the goal is to make it possible for the user to view the institution via Google Maps from the e-Doctor application. The user can also select the “Login” option from the “Institution” tab without having to return to the homepage. For each facility, there is a “View practices” option, marked with a blue button on the right, which allows users to view a list of practices within the selected facility. Furthermore, the user can view the doctor’s schedule within

the selected practice as shown in the image. The schedule is divided into three shifts: morning, afternoon and night. Each shift displays the name of the doctor working during that time, as well as the corresponding dates when they are on the schedule. At the bottom of the window, there is a “Close” button that allows the user to return to the previous page.



Fig. 4. e-Doctor logo

3.3 Patient

The user can log in to the system via the “Login” form. When the user is logged in to the system as a patient, they have the option to view their personal profile within the e-Doctor application. The options available in the top navigation bar are: Home, Institutions, Profile, Chat and Diet Plan. Within the application, the patient is currently in the “Profile” section, and the “Allergies” tab is active, which is marked in blue. Allergies contain the name of the allergy and an explanation for each allergy. On the left, the user can select other tabs within the profile, such as medical history, medications and findings. This view allows for a quick and easy overview of allergies and their descriptions. The patient has the option to view the medications they use and thus obtain information about them. Two medications with basic information are shown. In this way, the patient is provided with a quick overview of the medications they use with important information about each medication. Furthermore, when selecting the “Findings” tab, the user can view various medical findings. In further implementation, the patient will have the option to view a PDF version of the findings, while VirtualDoc will be able to explain the findings to the patient in detail, providing additional information and explanations about the results. The page displays various findings with basic information such as the date of the findings, the type of findings, and a brief description of the results. The “Chat” option allows communication with a virtual assistant called VirtualDoc within the e-Doctor application. The user asks a question related to their health, and VirtualDoc answers. This chat allows patients to obtain quick and simple information about medications, symptoms, and other medical topics via a virtual assistant, which contributes to improving the experience and facilitating access to information. In the further phase of the application development, the goal is to enable the uploading of a document based on which VirtualDoc can explain the finding. The diet plan was automatically generated using the ChatGPT API, which was passed all the relevant medical data of the patient. Based on this data,

ChatGPT created a personalized diet plan that takes into account the patient's health condition, specific dietary needs, and potential allergies. A meal plan is displayed for each day of the week, where a list of ingredients is listed for each meal that is tailored to the patient. The goal is to upgrade the existing diet plan display so that a detailed process of preparing a meal is shown. An additional upgrade would be to select ingredients that the patient does not prefer so that certain ingredients are not used when generating a weekly meal plan. This approach allows the patient to monitor and prepare meals that are in line with their health condition, thus encouraging a healthier lifestyle.

3.4 Doctor

The user can log in to the system via the “Login” form. When the user is logged in to the system as a doctor, they have the option to manage patient profiles within the e-Doctor application. The doctor can view detailed information about each patient, including their medical history, allergies, medications, and medical findings. This provides the doctor with a structured and clear overview of the patient's health status.

Within the application, the doctor can currently review existing medical findings and records. In the further phase of development, doctors will have the ability to upload new medical findings in PDF format, while AI DocAssist will be able to assist them by providing explanations and additional insights about the results. In this way, the doctor will receive support in interpreting findings and will be able to offer patients more precise and personalized healthcare.

In future implementation, this functionality will be upgraded so that doctors can adjust therapy dynamically and receive alerts about interactions or important changes in the patient's condition.

The “Chat” option enables communication with DocAssist, a virtual assistant integrated into the e-Doctor application. The doctor can ask questions related to patient health data or treatment options, and DocAssist provides quick and relevant answers with links to medical articles. This functionality supports doctors in their decision-making process, reduces the administrative burden, and helps in accessing information more efficiently.

Additionally, the doctor can view information about healthcare institutions. In future upgrades, this section will be expanded with details such as working hours and advanced coordination options, which will further improve the management of patient referrals and medical services.

4 Conclusions

Modern technology is still some way from achieving “third wave AI” capabilities, also known as “artificial general intelligence” or “strong AI”. These terms refer to the AI systems that approach super-intelligence, going beyond narrow tasks to approach other challenges in a way that approaches human or super-intelligence

[11]. During this research, the main goal was to show the application of artificial intelligence in today's time, taking into account the security of the business process. Cybersecurity is an arms race and the fight against cybercrime has moved inside the network. Data can be stolen, services interrupted, and systems destroyed once a network perimeter security is breached. Technological solutions to cyber threats dominate the landscape (e.g. security architecture, artificial intelligence – AI, machine learning). In order to develop dynamic, adaptive systems, researchers are using artificial intelligence models to simulate cyber war game scenarios. Drawing upon Game Theory and machine learning to balance defenders' advantage, work on modeling and simulation for artificial intelligence in adaptive defense systems is showing promise. Along with this research, in order to adequately develop AI capable of deceiving a human attacker, it is critical to evaluate these models with human subjects research. It is one thing to model a logical, binary situation according to a small set of rules. However, humans do not always operate in a rational manner. In order to understand how to model commonly displayed irrational decision-making, human subjects research is key. Investigating cognitive bias as a deceptive practice in defensive operations stands to inform future AI development [14]. The work is divided into theoretical and practical parts. The theoretical part provides general knowledge about artificial intelligence, its history, types and branches, with an emphasis on its application. The application of artificial intelligence has become indispensable in many areas, but its impact in healthcare is particularly highlighted, which forms the basis of this thesis. The practical part focuses on ChatGPT, an advanced language model developed by OpenAI. In the paper, ChatGPT was used as a virtual assistant in the e-Doctor application as a virtual assistant for providing health information and advice, which facilitates communication between patients and doctors. Through diagrams of use cases, activities and business processes, key processes within the application are visually presented. The development of the e-Doctor web application prototype represents an important step in the modernization of healthcare services, providing patients and healthcare professionals with faster and easier access to information. The prototype thus shows the potential for improving the quality of healthcare, reducing waiting for information and better managing patient data. In order to further improve the functionality and efficiency of the e-Doctor application, it is necessary to include the following activities in the next stages of development. The first activity involves integration with real health data by establishing an API connection to connect the application to the healthcare system database. As for security upgrades, it is necessary to additionally ensure data privacy, which is achieved by replacing the current authentication system with the National Identification and Authentication System NIAS and e-Authorization. In order to ensure GDPR (General Data Protection Regulation) standards, a regular review of security requirements is required. For users with certain disabilities such as blindness and low vision, efforts will be made to improve accessibility through increased font size, use of screen readers, high contrast, etc. For patients, it is planned to introduce the possibility of direct video consultations with doctors via the application, which will enable the

integration of video calling systems. The last activity includes conducting user experience testing to identify potential problems or improvements. Analyzing the data [11] also revealed a series of challenges facing governments. These include the need for governments to modernize procurement mechanisms for emerging technologies and adapt to the increasingly critical role of the private sector, how AI may threaten aspects of strategic stability, and the need for innovators and militaries to enhance testing, evaluation, verification and validation (TEVV) processes for AI-enabled systems. Summarizing the conclusions section we consistently present the main results obtained as follows. Motivation for the study, among others relevant challenges, was the idea of using AI to reduce administrative burdens and improve service delivery in public health as both practical and forward-looking. The main result achieved by this study is the development of a functional prototype of the e-Doctor application with the potential to create a functional application after testing and a positive outcome of the acceptance test. This is achieved by fully satisfying non-functional (system) requirements and functional user requirements, before the moment of delivery. The aforementioned limitations arise from the fact that this is a dynamic system subject to further development, refinement and changes arising from (indicated) business needs and user requirements. The scientific contribution of this paper is recognized as timely and relevant contribution by exploring the integration of ChatGPT into a public healthcare system through the development of a prototype web application e-Doctor. The study aims to improve citizens' access to medical information via the Croatian e-Citizens platform which presents theoretical overview of AI technologies with a detailed account of the design and functionality of the prototype. The focus of the study was on real-world implementation as we clearly described the structure and features of the application, including its use of PHP, MySQL, and Symfony, as well as the integration of ChatGPT for patient interaction. The user experience, especially for patients, is thought out – including access to medical history, medications, and personalized diet plans generated by ChatGPT. The design seems user-friendly and reflects genuine concern for accessibility. Therefore, the scientific contribution of the study is also reflected in the application of the foundations of AI through applied research in the field of applied informatics, by defining the goals, processes and resources necessary for implementation, and the successful implementation itself, especially taking into account meeting cybersecurity requirements as a fundamental prerequisite for the functioning and survival of modern application systems. The limitations of the study are observed as follows. The application is designed to function in a modern environment marked by disinformation and misinformation along with AI-generated responses. Therefore it is necessary to assess the risks of aforementioned information misuse by the adversary as well as risks of overreliance from the users' perspective. User privacy and data security are ensured by applying cyber security settings, relying on cyber defense processes within the business system in the same time. But during integration with national health databases all conditions arising from the legal framework of healthcare must also be met. Finally, legal, ethical and societal questions and conditions regarding the

existence of mechanisms for accountability, if the AI gives incorrect advice, are also raised. According to our best knowledge, these questions are subject to ongoing consideration and with the development and application of AI solutions in everyday business there is a growth in their multiplication and complexity, which brings opportunity for further challenging research.

Acknowledgments. This work was supported by the project 030TUKE-4/2023 – “Application of new principles in the education of IT specialists in the field of formal languages and compilers”, granted by the Cultural and Education Grant Agency of the Slovak Ministry of Education.

References

1. How does Chat GPT work? – ATRIA Innovation. <https://atriainnovation.com/en/blog/how-does-chat-gpt-work/>. Accessed 08 May 2024
2. What is the history of artificial intelligence (AI)? | tableau. <https://www.tableau.com/data-insights/ai/history#ai-boom>. Accessed 02 May 2024
3. Ayers, J.W., et al.: Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**(6), 589–596 (2023). <https://doi.org/10.1001/JAMAINTERNMED.2023.1838>
4. Betz, S.: 7 types of artificial intelligence | built in. <https://builtin.com/artificial-intelligence/types-of-artificial-intelligence>. Accessed 02 May 2024
5. Bohnenberger, T.: AI & The Industry. https://www.google.hr/books/edition/AI_The_Industry/uzboEAAAQBAJ. Accessed 02 May 2024
6. Bowman, J.: How is AI used in healthcare? | the motley fool. <https://www.fool.com/investing/stock-market/market-sectors/information-technology/ai-stocks/ai-in-healthcare/>. Accessed 06 May 2024
7. Briganti, G., Moine, O.L.: Artificial intelligence in medicine: today and tomorrow. *Front. Med. (Lausanne)* **7** (2020). <https://doi.org/10.3389/fmed.2020.00027>
8. Copeland, B.J.: Artificial intelligence - machine learning, robotics, algorithms | britannica. <https://www.britannica.com/technology/artificial-intelligence/Is-artificial-general-intelligence-AGI-possible>. Accessed 01 May 2024
9. Das, S.: What are the components of AI? - ellow.io. <https://ellow.io/components-of-ai/>. Accessed 02 May 2024
10. Digital Adoption Team: The most important branches of artificial intelligence you need to know. <https://www.digital-adoption.com/branches-of-artificial-intelligence/>. Accessed 02 May 2024
11. Ertan, A.: Exploring the security implications of artificial intelligence in military contexts. Ph. D. thesis, Royal Holloway, University of London (2022). <https://pure.royalholloway.ac.uk/ws/portalfiles/portal/46513266/2022ErtanAPhD.pdf>
12. Gulati, A.: The history of artificial intelligence (AI): a detailed guide. [https://www.knowledgehut.com/blog/data-science/history-of-ai#the-second-ai-winter-\(1987-1993\)](https://www.knowledgehut.com/blog/data-science/history-of-ai#the-second-ai-winter-(1987-1993)). Accessed 02 May 2024
13. Ingram, A.: ChatGPT. Greenbooks Editore. <https://www.google.hr/books/edition/ChatGPT/cGrfEAAAQBAJ>. Accessed 08 May 2024

14. Johnson, C.K., Tassel, R.W.V., Rogers, A., Shade, T., Ferguson, K.: Adversarial cognitive engineering (ace) and defensive cybersecurity: Leveraging attacker decision-making heuristics in a cybersecurity task. In: Proceedings of the 57th Hawaii International Conference on System Sciences (2024)
15. Kumble, G.P.: Practical Artificial Intelligence and Blockchain: A Guide To Converging Blockchain and AI to Build Smart Applications for New Economies. Packt Publishing Ltd. (2020)
16. Malik, E.: Artificial intelligence (AI) and ChatGPT: history and timelines. <https://www.officetimeline.com/blog/artificial-intelligence-ai-and-chatgpt-history-and-timelines>. Accessed 08 May 2024
17. Marr, B.: How is AI used in healthcare - 5 powerful real-world examples that show the latest advances. <https://www.forbes.com/sites/bernardmarr/2018/07/27/how-is-ai-used-in-healthcare-5-powerful-real-world-examples-that-show-the-latest-advances/>. Accessed 06 May 2024
18. Mijwil, M.M.: History of artificial intelligence **3** (2015). <https://doi.org/10.13140/RG.2.2.16418.15046>
19. Pandey, V.: Artificial Intelligence for Students: A Comprehensive Overview of AI's Foundation, Applicability, and Innovation (English Edition). Bpb Publications (2024). <https://books.google.hr/books?id=ptq1EAAAQBAJ>. Accessed 01 May 2024
20. Roser, M.: The brief history of artificial intelligence: the world has changed fast — what might be next? - our world in data. <https://ourworldindata.org/brief-history-of-ai#article-citation>, Accessed 02 May 2024
21. Sung, J.J., Stewart, C.: Artificial Intelligence in Medicine. Elsevier. https://www.google.hr/books/edition/Artificial_Intelligence_in_Medicine/gE7VEAAAQBAJ. Accessed 03 May 2024
22. The Technology Edits: ChatGPT: Potential Useases (2023). https://www.google.hr/books/edition/ChatGPT_Potential_Usecases/sxWtEAAAQBAJ. Accessed 04 June 2024
23. TUTOO: ChatGPT for Language Teachers. Tredition (2024)

Author Index

A

Alaminos, David 227
Almahameed, Ala Ali 79
Almeida, António 69
Almeida, Inês 69
Anwar Lashari, Tahira 516
Arias-Oliva, Mario 79, 274
Ascama, Héctor Dave Orrillo 476, 566

B

Bagheri, Sally 314
Bassi, Eleonora 204
Blasco López, María Francisca 27
Bolarín-Miró, Elisabet 60
Bracamonte, Vanessa 428
Brandao, Pedro 589
Brooks, Laurence 451

C

Cabeças, António 385
Calzati, Stefano 48
Camacho-Ruiz, Miguel 252
Canorea-García, Rafael 412
Carmona Martínez, María Mercedes 60, 252
Carrasco, Ramón-Alberto 3, 27
Carrasco-Aguilar, Álvaro 227, 252
Coelho, Lina 116
Coelho, Luana 420
Correia, Fernando 566
Cruvinel Júnior, Laercio 192, 385, 476, 566

D

de Andrés Sánchez, Jorge 274
Debas, Felipe 420
Dediu, Adrian-Horia 105
Doore, Stacy A. 326
Duarte, Fernanda 263
Duda, Ewa 289, 338

E

Eneman, Marie 504

F

Fernández-Avilés, Gema 3
Fleischman, William 490
Frasquilho, Afonso 566
Fukami, Yoshiaki 464
Fukuta, Yasunori 397, 428

G

Galinec, Darko 599
Gerdes, Anne 350, 360
Giouvanopoulou, Konstantina 15
Guillén-Pujadas, Miguel 227

H

Héctor Ascama, Orrillo 192
Heimo, Olli I. 215
Hori, Masashi 397
Hosell, Sarah 153
Houghton Torralba, Miguel 27

I

Ilijoski, Bojan 169
Isobe, Taichi 397
Isohara, Takamasa 428

J

Jabeen, Farzana 516

K

Källström, Klara 441
Kampmann Walther, Bo 360
Karkaletsis, Vangelis 15
Karpouzis, Kostas 15
Karvonen, Antero 371
Kesar, Shalini 579
Kirkpatrick, Michael S. 239
Kubanek, Mariusz 531, 542

L

Lehtonen, Teijo 215
 Leikas, Jaana 371
 Ljungberg, Jan 504
 Lopes Fernandes, Adriana 385
 Lopes Fernandes, Guilherme 566
 López-Arranz, Asunción 60
 López-Galiacho Perona, Juan Luis 79
 Lourenço, Nuno 129

M

Marques da Silva, Mário 192, 476, 566
 Marques, Camila 129
 Martínez Hernández, Rosalba 35
 Matos, Joana M. 105
 Meyerson, Denise 69
 Mikhaylova, Anjela 451
 Młynarczyk-Karabin, Ewelina 289, 338
 Monteiro, Lúcia 412
 Monteiro, Rosa 116, 129
 Montero-Lorenzo, José-María 3
 Murata, Kiyoshi 397, 428

N

Núñez-Guerra, Paula M. 95

O

Ollé Sesé, Manuel 274
 Orita, Akiko 301
 Orito, Yohko 397, 428

P

Pagallo, Ugo 204
 Pais D'Aguiar, Filipa 263
 Papademas, Michael 15
 Parra Meroño, María Concepción 252
 Pelegrín-Borondo, Jorge 79
 Pérez-Portabella, Antonio 274
 Pöntinen, Sanni 371

Q

Qayyum, Ahmed 326

R

Radaković, Davorka 141
 Rashid, Mehvish 516
 Rosenbloom, Leah Namisa 490
 Rua, Orlando Lima 412, 420

S

Sai, Hidenobu 397
 Sako, Kazue 464
 Santos, Mariana 116
 Shu, Ziwei 27
 Siedlecka-Lamch, Olga 554
 Sigfrids, Anton 371
 Silva, Nuno 263
 Souto-Romero, Mar 227, 412
 Stahl, Bernd 451
 Steingartner, William 141, 599
 Sultana, Sidra 516
 Suzuki, Shizuka 397
 Szymoniak, Sabina 531, 542, 579

T

Torres-Coronas, Teresa 181
 Treviño Serna, Melissa Eugenia 35

V

Velasco-López, Jorge-Eusebio 3
 Viana, Tomás 566

W

Wessberg, Nina 371
 Więckiewicz-Modrzewska, Julia 289, 338
 Wilford, Sara 451

Y

Yamamoto, Tomonori 397
 Yáñez Luna, Juan Carlos 35

Z

Zdravkova, Katerina 169
 Zeko, Natalija 599
 Ziouvelou, Xenia 15