ISABEL ALVAREZ

ADRIAN-HORIA DEDIU

COORDINATION

ETHICOMP 2025

22ND INTERNATIONAL CONFERENCE ON THE ETHICAL AND SOCIAL IMPACTS OF ICT LISBON

CONFERENCE COMPANION VOLUME

AUTÓNOMA
CULTURA
edições

ISABEL ALVAREZ ADRIAN DEDIU
COORDINATION

ETHICOMP 2025

22ND INTERNATIONAL CONFERENCE
ON THE ETHICAL AND SOCIAL IMPACTS OF ICT LISBON

EXTENDED ABSTRACTS

AUTONOMA
CULTURA

DATA SHEET

TITLE
ETHICOMP 2025: Conference Companion Volume

COORDINATION
Isabel Alvarez
Adrian-Horia Dediu

PUBLISHER
Universidade Autónoma de Lisboa | Autónoma Edições

EDITORIAL COORDINATION
Raquel Medina Cabeças

REVISION
Isabel Alvarez
Adrian-Horia Dediu
Autónoma Edições

DESIGN & PAGINATION
Raquel Medina Cabeças

ISBN 978-989-9002-62-3
DOI https://doi.org/10.26619/978-989-9002-62-3

The Cooperativa de Ensino Universitário, the founding body of the Autonomous University of Lisbon, promotes scientific production in various cultural segments, valuing the relationship between the academic community and society. In this way, it supports the edition of this publication, contributing to the dissemination of knowledge.

Table of Contents

OPEN TRACK

Let's Re-evaluate the Best Practices of Cybersecurity

Gender and Future Digital Technologies

The normative challenges of AI literacy

Artificial Intelligence: Interpretability and Measurement of its Underlying Ethics

VALUES: THE FOUNDATION FOR RESPONSIBLE TECHNOLOGY AND DIGITAL INTERACTIONS

HOW IS AI CHANGING US? HOW DOES IT CHANGE THE WORLD?

ETHICAL CHALLENGES FOR BUSINESS IN A DIGITAL WORLD

AI, SOCIAL NETWORKS, AND THEIR INFLUENCE ON YOUTH

PREFACE

This volume contains the collection of the extended abstracts submitted to ETHICOMP 2025: 22nd International Conference on the Ethical and Social Impacts of ICT, held on September 17- 19, 2025, in Lisbon, Portugal.

Conference Overview

ETHICOMP is a well-established international conference series that explores the ethical and social implications of information and communication technologies (ICTs). The conference continues a longstanding tradition of fostering interdisciplinary dialogue and is held approximately every 18 months. Previous editions have been held in cities such as Logrono (2024), Turku (2022), Bologna (2020), and Paris (2018), among others. Each conference has contributed to the evolving discourse on the societal impacts of digital innovation.

ETHICOMP welcomes contributions from scholars at all career stages. In particular, the conference encourages student participation and provides an open platform for emerging voices in the field. Accordingly, this volume includes extended abstracts, enabling early-stage work and ongoing research to be shared within the community. A separate volume containing the full papers accompanies the conference proceedings.

Tracks and Themes

ETHICOMP 2025 is a multitrack conference that brings together diverse sessions at the intersection of ethics, technology, society, and education. Designed to foster interdisciplinary dialogue, the program offers both thematic depth and crossdomain perspectives. Domain experts chair each track, reflecting the breadth of ethical challenges in contemporary information and communication technology (ICT)—from algorithmic fairness and data protection to AI governance, digital pedagogy, and cyberethics. To reflect the diverse ethical questions raised by digital technologies, ETHICOMP 2025 is organized around the following topical areas:

Let's Reevaluate Cybersecurity Best Practices

Explores how technologies like AI, IoT, and machine learning reshape cybersecurity norms, calling for ethical frameworks to address threats, privacy, and digital responsibility.

Ethical Challenges of Human Relations in Education in the AI Era

Investigates how AI alters teacher–student relationships, peer interaction, and emotional development in digital learning environments, raising questions of equity and human agency.

Gender and Future Digital Technologies

Examines how emerging technologies reflect, reinforce, or challenge gender norms and disparities across design, policy, and access to innovation.

The Normative Challenges of AI Literacy

Addresses the ethical, legal, and educational implications of AI literacy, highlighting the risks of exclusion and the need for normative guidelines in the responsible use of AI.

(De)stigmatizing Digital Interactions and Technologies

Analyzes how online platforms and AI systems both perpetuate and combat stigma, especially in contexts of mental health, sexuality, and marginalized identities.

Artificial Intelligence: Interpretability and Measurement of its Underlying Ethics

Focuses on transparency and ethical behavior in autonomous systems, emphasizing explainable AI (XAI), accountability, and trust in algorithmic decisions.

Values: The Basis for Responsible Technology and Digital Interactions

Explores how ethical values can be embedded throughout the lifecycle of ICT systems, with attention to design trade-offs and case studies in cybersecurity and education.

How is AI Changing Us? How Does it Change the World?

Reflects on how AI and generative technologies reshape behavior, agency, norms, and institutions, inviting perspectives on both positive and disruptive outcomes.

Ethical Challenges for Business in a Digital World

Considers ethical decision-making in digitally transformed organizations, from reputation and governance to empirical studies of ethical strategy.

AI, Social Networks, and Their Influence on Youth

Examines how AI-powered social media affects the well-being, privacy, and civic engagement of young people, and explores the ethical safeguards necessary for youth-centric platforms.

Open Track Welcomes interdisciplinary submissions that explore emerging or unconventional areas in computing ethics not addressed in the listed topics.

Organization

We are grateful for the help and support received from the organizing institution: – Universidade Autónoma de Lisboa. Besides our university, the conference received help from various collaborators from the following institutions:

- Universidad Complutense de Madrid
- Universidad Rovira i Virgili, Tarragona

- Universidad de Barcelona
- Instituto Politécnico de Tomar
- Centro de Investigação em Cidades Inteligente
- Centro de Investigação Autonoma TechLab
- Center for Computing and Social Responsibilities, DeMontfort University, Leicester
- Instituto Politécnico do Porto
- COMEGI – Universidades Lusiada de Lisboa
- Meiji University, Tokyo, Japan
- Southern Utah University

Below is our organizing leadership and committee:

Organizing Committee Chairs:

- Reginaldo de Almeida, Vice President, Universidade Aut´onoma de Lisboa
- Mário Marques da Silva, Director, Faculty of ICT, Universidade Aut´onoma de Lisboa

Organizing Committee Members:

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

Submission and Review

A total of 109 submissions were received for this edition. Each submission was evaluated by at least two reviewers, with an average of 2.2 members of the Program Committee or external reviewers per paper. After a thorough deliberation, the committee accepted 98 papers for presentation during the conference.

Acknowledgements: We extend our sincere gratitude to the organizers, but especially to all the authors who showed interest in our conference.

May 23, 2025
Lisbon
Isabel Alvarez
Raquel Medina Cabeças
Adrian Horia Dediu
Nuno Silva

SOCIAL RISKS OF BRAIN-MACHINE INTERFACE USAGE: QUESTIONNAIRE SURVEY OF PEOPLE WITH AND WITHOUT DISABILITIES

Yohko Orito[1][0000-0002-4293-1722], Tomonori Yamamoto[2][0000-0002-9600-3876], Shizuka Suzuki[3][0009-0000-6868-1306], Hidenobu Sai[4][0000-0002-6010-4747], Kiyoshi Murata[5][0000-0002-5632-8078], Yasunori Fukuta[6][0000-0002-2712-5538], Taichi Isobe[7][0000-0002-2477-0830], Masahi Hori[8][0000-0002-0651-1544]

Keywords: Brain machine interface, support for people with disabilities, cyborgisation, questionnaire survey

Extended Abstract

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, particularly 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 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

1Ehime University, Matsuyama, Ehime 790-8577, Japan.
2 Ehime University, Matsuyama, Ehime 790-8577, Japan.
3 Ehime University, Matsuyama, Ehime 790-8577, Japan.
4 Ehime University, Matsuyama, Ehime 790-8577, Japan.
5 Meiji University, Chiyoda, Tokyo 101-8301, Japan.
6 Meiji University, Chiyoda, Tokyo 101-8301, Japan.
7 Health Sciences University of Hokkaido, Ishikari-gun, Hokkaido 061-0293, Japan.
8 Waseda University, Shinjuku, Tokyo 169-8050, Japan, Orito.yohko.mm@ehime-u.ac.jp.

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 and ethical considerations of cyborg technologies including BMIs. It was also intended to enable a comparison of opinions and recognitions between the two groups. For people with and without disabilities, the same question items were administered as much as possible. The questionnaire survey with people with and without disabilities 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 848responses, 683 were deemed valid. The gender and age distributions of the respondents are listed in Table 1.

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. As summarised in the tables, a slight demographic bias in terms of gender and age can be 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						
	10s	20s	30s	40s	50s	60s (60-65)	**Total**
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						
	10s	20s	30s	40s	50s	60s	Total
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 allowed in terms of symptoms) (n = 87)

TYPES / SYMPTOMS (Multiple Answer)	Congenital disabilities	Acquired disabilities	Both (congenital/ acquired)	No answer	Total
Physical limb disability	39	43	1	1	84
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

Among 770 valid responses from the both groups, 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 these questions (p = .5260)

On the other hand, regarding the recognition of the expected benefits, possibilities and risks of BMIs, some significant differences and similarities were observed in responses between people with and without disabilities. Based on the survey results, this study attempted 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.

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), Mr. Yukio Takeda, Dr. Yoshitaka Moritsugu, 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 Computational Biology **15**(1), e1006633 (2019)
2. Chaudhary, U., Birbaumer, N., Curado, M. R.: Brain-machine interface (BMI) in Paralysis. Annals of physical and rehabilitation medicine **58**(1), 9–13 (2015)
3. Liu, M., Ushiba, J.: Brain–machine Interface (BMI)-based neurorehabilitation for post-stroke upper limb paralysis. The Keio Journal of Medicine **71**(4) 82–92 (2022)
4. Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T., Hori, M.: Is a brain machine interface useful for people with disabilities? Cases of spinal muscular atrophy. Proceedings of ETHICOMP 2024: The Leading Role of Smart Ethics in the Digital World, pp.9–19. (2024)
5. Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T., Hori, M.: The social implications of brain machine interfaces for people with disabilities: Experimental and semi-structured interview surveys. 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)
6. Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T., Hori, M.: How a brain-machine interface can be helpful for people with disabilities? Views from social welfare professionals. In Mario Arias Oliva, Jorge Pelegrín Borondo, Kiyoshi Murata and Ana María Lara Palma (eds.), Moving Technology Ethics at the Forefront of Society, Organisations and Governments, pp.103–115. (2021)
7. Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T., Hori, M.: 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. Logroño, Spain: Universidad de La Rioja (2020)
8. Cabinet Office, White paper on persons with disabilities, https://www8.cao.go.jp/shougai/whitepaper/r06hakusho/zenbun/siryo_01.html (in Japanese), last accessed 2024/12/31.

A TALE OF TWO CATASTROPHES THE MACHINE STOPS &THE PROPAGANDA-INDUSTRIAL COMPLEX GOES CRITICAL

William Fleischman[1], Leah Namisa Rosenbloom[2]

Keywords: technology ethics · social media · propaganda-industrial complex · E.M. Forster · abolition technology · facism · imperialist White supremacist capitalist cisheteropatriarchy.

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 in the past two decades culminating in the resurrection of a tech billionaire-backed authoritarian to one of the most powerful executive offices in contemporary world government. Forster's premonitory intuition concerning the fragility of massive and complex technological systems provides a rich retrospective on problems in computer and information ethics.

The symbiotic alliance between billionaire social media owners and authoritarians, which Rosenbloom and Fleischman named the propaganda-industrial complex (PropIC) [19], has escalated over the past several years. Platforms have moved from a performative and exploitative treatment of content moderation, through the cynical abandonment of content moderation, towards outright support for fascist White supremacist capitalist cisheteropatriarchy. This pivot has openly and explicitly cemented platform support for authoritarian propaganda.

We discuss *The Machine Stops* and its place in the curriculum of computer ethics. We identify a connection between The Machine in Forster's story and contemporary social media platforms that forms a bridge to the second IRL catastrophe. We analyze the current state of the PropIC and propose a pivot of our own: from ineffective top-down regulatory interventions to human-scale abolitionist interventions, which can grow resistance and resilience to The Machine from the bottom up.

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* [8].

1 Villanova University, Villanova, PA 19085, USA, william.fleischman@villanova.edu
2 Northeastern University, Boston, MA 02115, l.rosenbloom@northeastern.edu

The inclusion of the case of the Therac-25 hardly requires comment [6]. 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.

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.

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.

La Vie Moderne: Assimilating the Machine to Social Media

Contemporary students readily associate The Machine with a social media platform, most typically Facebook. Facetime is the obvious modern referent for the of iconic "blue plate" through which Kuno and Vashti communicate. This identification functions like 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 untrammeled ability to communicate with others creates a perceptible pressure in the rhythm of life that predisposes to "irritation – a growing quality in that accelerated age" [8].

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. 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" [25].

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.

What Forster Could Not Foresee

The idea that mutinous individuals motivated by the thirst for power and dominance over others might venture 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 his "meditation."

The Propaganda-Industrial Complex

The Machine undergirding the propaganda-industrial complex (propIC) [19], 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 to unmasked support for fascist White supremacist capitalist cisheteropatriarchy [10,12,4,9]. 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 [2,15,24]. 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 farright content. Zuckerberg's gratuitous persecution of trans Meta employees [20], Musk's glorification of Naziism [23], and both billionaires' conspicuous placement on the U.S. presidential inauguration stage suggest that the technologies undergirding the propIC are, at the root, being operationalized to advance systems which encourage genocide 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 [5]—technological weapons of the military- and prison-industrial complexes rather than public resources. They must be treated as such.

Studies of content moderation [13,7,14,17] 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, is willing to platform users who contribute to the ongoing systematic erasure and attempted genocide of trans and LGBTQ+ people [16]. We have come to believe that the problem is more radical than just trying to muck out the Augean stables of false, tendentious and hateful material online. On our view, the combination of network effects and the untrammeled, 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." [21]

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. Our bodies can speak truth to power.

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 [3,22,18, 11] 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.

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. 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)
3. Benjamin, R.: Race after technology: Abolitionist tools for the new jim code. SocialForces (2019)
4. Benjamin, R.: Viral Justice: How We Grow the World We Want. Princeton University Press (2022)

5. Dedezade, E.: Tiktok users report anti-trump content being hidden following platform's unbanning. Forbes (jan 2025), https://www.forbes.com/sites/esatdedezade/2025/01/22/tiktok-users-reportanti-trump-content-being-hidden-following-platforms-unbanning/
6. Fleischman, W., Crawford, J.: Once again, we need to ask, "what have we learnedfrom hard experience?". In: Societal Challenges in the Smart Society. pp. 561–572. Universidad de La Rioja (2020)
7. Fleischman, W.M., Langan, N., Rosenbloom, L.N.: Endless forms most deceitful. ETHICOMP 2022 p. 104 (2022)
8. Forster, E.M.: The Machine Stops. William Heinemann (1909)
9. hooks, b.: Yearning: Race, Gender and Cultural Politics. South End Press (1990)
10. hooks, b.: Teaching to Transgress: Education as the Practice of Freedom. Routledge(1994)
11. Jones, S.T., melo, n.a.: We tell these stories to survive: Towards abolition in computer science education. Canadian Journal of Science, Mathematics and Technology Education 21(2), 290–308 (2021)
12. King Jr, M.L.: Letter from birmingham city jail. American Friends Service Committee (1963)
13. Marantz, A.: Why facebook can't fix itself. The New Yorker 12 (2020)
14. Newton, C.: The trauma floor: The secret lives of facebook moderators in america.The Verge (2019), https://www.theverge.com/2019/2/25/18229714/cognizantfacebook-content-moderator-interviews-trauma-working-conditions-arizona, accessed: 2025-04-04
15. Ortutay, B., AP: Meta rolling back hate speech rules: Zuckerberg's recentelections catalyst. Fortune (2025), https://fortune.com/2025/01/09/meta-rolling-back-hate-speech-rules-zuckerberg-recent-elections-catalyst/
16. 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-petitionto-ban-jesse-singal-over-anti-trans-views-harassment/, accessed: 2025-04-04
17. Perrigo, B.: Facebook content moderators in kenya file lawsuit over abuse, working conditions. Time (2021), https://time.com/6175026/facebook-sama-kenyalawsuit/
18. Rosenbloom, L.N.: Cryptography and collective power. Cryptology ePrint Archive(2024)
19. Rosenbloom, L.N., Fleischman, W.: Social media and the rise of the propagandaindustrial complex. In: ETHICOMP 2021: Moving technology ethics at the forefront of society, organisations and governments. pp. 503–510. Universidad de La Rioja (2021)
20. Silva, C.: Meta employees revolt after zuckerberg removes tampons fromworkplace. Mashable (jan 2025), https://mashable.com/article/mark-zuckerbergremove-tampons-meta-employees-revolt
21. Snyder, T.: On Tyranny: Twenty Lessons from the Twentieth Century. Tim DugganBooks: New York (2017)
22. Strong, L., Das, A.: Abolition science mission. https://www.abolitionscience.org/(2018)
23. Treisman, R.: Elon musk's relationship with german far-right afd under scrutinyfollowing holocaust comments. NPR (2025), https://www.npr.org/2025/01/27/nxs1-5276084/elon-musk-german-far-right-afd-holocaust
24. VergeStaff: The fallout of meta's content moderation overhaul. The Verge(2025), https://www.theverge.com/24339131/meta-content-moderation-factcheck-zuckerberg-texas
25. Weizenbaum, J.: On the impact of the computer on society: how does one insult amachine? Science 176(4035), 609–614 (1972)

INFORMATION ETHICS EDUCATION FOR BUSINESS PEOPLE

KIYOSHI MURATA[1] [0000-0002-5632-8078], YOHKO ORITO[2] [0000-0002-4293-1722], YASUNORI FUKUTA[3] [0000-0002-2712-5538], ANDREW A. ADAMS[4] [0000-0003-0410-802X], TATSUYA YAMAZAKI[5] [0000-0001-8506-1731]

Keywords: Information Ethics, Education, Business People.

Extended Abstract

This study addresses how information ethics education for business people should be established. Despite the fact that a large majority of computing activities — the development, operation and use of information and communication technology (ICT)-based information systems — are carried out by business organisations, an effective and well-organised educational programme of information ethics for business people, including engineers, non-technical users and administrators, has not been developed and implemented, whereas there are many attempts to propose proper computer ethics education programmes for computer engineers and computer science and engineering students (see, e.g., [1-3]). Worse still, business people seem not to be well aware of the meaning and social significance of ethical initiatives in a business setting. This means that an information ethics education programme for business people would not work well unless it had been developed taking such a reality into consideration.

In Japan, since the end of 1990s, the practices of business ethics including "compliance" and "corporate social responsibility (CSR)" have been engaged in by many companies, centred mainly on the efforts of large corporations to avoid public criticism about their business activities. Recently, reflecting the increasing social interest in environmental sustainability, "environmental, social and corporate governance (ESG)" and "the sustainable development goals (SDGs)" have been added to the agenda of business ethics. It is often argued by business consultants and news media that companies' approaches to these are important as a response to those investors who are sensitive to environmental concerns. "Ethical, legal and social issues/implications (ELSI)" and "responsible research and innovation (RRI)" are increasingly emphasised as issues that organisations involved in technology development and deployment should address to avoid blame for them. On the other hand, the boundary of business ethics has blurred as various issues like diversity and inclusion in the workplace and data protection have

1 MEIJI UNIVERSITY, CHIYODA, TOKYO 101-8301, JAPAN

2 EHIME UNIVERSITY, MATSUYAMA, EHIME

3 MEIJI UNIVERSITY, CHIYODA, TOKYO 101-8301, JAPAN

4 MEIJI UNIVERSITY, CHIYODA, TOKYO 101-8301, JAPAN

5 UNIVERSITY OF TOYAMA, TOYAMA, TOYAMA 930-8555, JAPAN, KMURATA@MEIJI.AC.JP

come to be considered business ethics issues. Actually, in Japan, many companies have depended on "business ethics solution packages" that consulting firms provide them to address such issues, demonstrating their vague or trivialised understandings of ethics in a business context. Nonetheless, most people don't cast doubt on the significance of business ethics for their companies' reputation management and profitability. However, if business people believe that they should act based on enlightened self-interest or that corporate ethical behaviour is necessary for ensuring (at least long-term) profitability of their companies through responding to social expectations and demands (regarding a logic of this sort, see, e.g., [4-9]), they may consider that ethics is not an end but a means in business, and business ethics would no longer deserve to be called ethics. If this is the case, it is hard to expect that business organisations, which play a pivotal role in the development, operation and use of ICT-based information systems, would proactively address issues related to information ethics and maintain responsible attitudes and behaviours towards computing. In reality, there are many ICT companies which maintain the attitude of "innovative first, consider consequences afterwards". What features should an information ethics education programme for business people have, then? This study attempts to get a clue to the answer to this question based on the authors' background study [10] and through interviews with Japanese university professors who teach "business ethics" and discussions with the audience at the ETHICOMP conference.

In the background study, a questionnaire-based survey of three respondent groups of Japanese people — (a) professional engineers, who were certified professional engineers working for companies; (b) ordinary employees, who were full-time, non-executive company employees without any professional qualifications; and (c) business students, who were 3rd- and 4th-year undergraduate students majoring in business — was conducted in July and August 2023 to investigate relevant Japanese people's attitudes towards business ethics. The results of the survey show that:

(a) Professional engineers were statistically more aware of each of the business ethics-related initiatives, except ELSI and RRI (which no group understood well), than business students and ordinary employees. This presumably owed much to continuing professional development programmes including engineering ethics instruction courses they were obliged to participate in to gain, retain and improve their knowledge, abilities and skills as professionals.
(b) Ordinary employees equivalently understood compliance and business ethics to business students, whereas business students understood CSR, ESG and the SDGs better than ordinary employees. The educational level of ordinary employee respondents was found to affect the level of their awareness of business ethics-related initiatives. They tended to underestimate the importance of each business ethics-related initiative compared to professional engineers and business students and be devoid of understanding of the rationale for ethics in a business setting. They lacked a sense of ownership over ethical issues in a business setting.
(c) There was no statistically significant difference in the awareness of ELSI and RRI among the three samples. The majority of respondents in each sample had not heard these terms, even though many current business innovation activities are technology-dependent.

(d) Most of the respondents in each sample considered that a company should address business ethics-related initiatives only as much as dealing with such initiatives did not worsen the company's long-term profit structure. Meanwhile, ordinary employees and business students tended to emphasise securing short-term profit rather than business ethics-related initiatives, whereas a majority of professional engineers expressed the opposite view.

These suggest that the development and implementation of a suitable information ethics education programme for ordinary business employees is an urgent necessity. Given that ordinary employees have already become computing practitioners thanks to the advancement of user-friendly applications including AI-based ones, and their lack of awareness of business ethics, effective in-house information ethics education for them has to be provided to make them sensitive to existing or potential ethical issues accompanied by the development, operation and/or use of ICT-based information systems, and to let them have sufficient knowledge about and a correct understanding of the meaning of information ethics in a business setting. This education should also cultivate their capacities and skills to consider and act ethically in a business setting as an organisational member and simultaneously as a citizen. A case method approach may be helpful for them to acquire these capacities, if there is a good mentor [11]. Because declarative and procedural knowledge necessary for perceiving and considering ethical issues related to computing should be updated on a constant basis, continuing education opportunities have to be provided. Ethical principles may have to be mentioned as part of the knowledge. However, it should also be taught that those principles are like a perfume: they should not be swallowed, but just be smelled. Any kind of indoctrination has to be avoided, and it is desirable that learners are encouraged to consider ethical problems autonomously, and as much as possible with a critical spirit.

The educational content should flexibly and appropriately be set according to the knowledge level of the target audience. Sufficient on- and off-the-job training opportunities for learning information ethics need to be provided to them. Moreover, professional engineers should be expected to play a pivotal role in fostering an ethical mindset among ordinary employees in their companies.

Business organisations have to establish necessary institutions within organisations to encourage and support employees to consider and behave ethically, such as the appointment of a chief ethics officer who is in charge of developing an ethical corporate culture and setting up well-organised information ethics education programmes. In terms of development and operation of information systems, building and operating the ICT governance system of "DevEthicsOps"[31] — the unification of creating and deployment of ICT systems using ethical practices — should be addressed beyond Ethics by Design [12], which can give the misconception that ethical approaches to technology are completed at the time of design. Such a governance system would be helpful in preventing the impact of threatening forces on ethical organisations [13].

3 This term was suggested by Dr Neil McBride of the Centre for Computing and Social Responsibility, De Montfort University, Leicester, UK, during a discussion with him.

Acknowledgments: This study was supported by the JSPS Grant-in-Aid for Scientific Research (C) 23K01545.

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

References

1. Reich, R., Sahami, M., Weinstein, J. M. and Cohen, H.: Teaching computer ethics: A deeply multidisciplinary approach. In: SIGCSE '20: Proceedings of the 51st ACM Technical Symposium on Computer Science Education, pp. 296–302. Association for Computing Machinery, New York (2020)
2. Stavrakakis, I., Gordon, D., Tierney, B., Becevel, A., Murphy, E., Dodig-Crnkovic, G., Dobrin, R., Schiaffonati, V., Pereira, C., Tikhonenko, S., Paul Gibson, J., Maag, S., Agresta, F., Curley, A., Collins, M. and O'Sullivan, D.: The teaching of computer ethics on computer science and related degree programmes. A European survey. International Journal of Ethics Education 7, 101–129 (2021)
3. Brown, N., Xie, B., Sarder, E, Fiesler, C. and Wiese, E. S.: Teaching ethics in computing: A systematic literature review of ACM computer science education publications. ACM Transactions on Computing Education 24(1), 1–36 (2024)
4. Porter, M. E. and Kramer, M. R.: The competitive advantage of corporate philanthropy. Harvard Business Review 80(12), 78–92 (2002)
5. Porter, M. E. and Kramer, M. R.: Strategy and society: The link between competitive advantage and corporate social responsibility. Harvard Business Review 84(12), 73–75 (2006)
6. Tanimoto, K.: Corporate social responsibility in the new era. In Tanimoto, K. (ed.), Corporate Social Responsibility: CSR and Stakeholders, pp. 2–34. Chuokeizai-sha, Tokyo (2004) (in Japanese)
7. Mizuo, J.: Strategic CSR management systems and SMIX21. In Mizuo, J. and Tanaka, H. (eds.), CSR Management, pp. 17–33. Japan Productivity Center Publishing, Tokyo, (2004) (in Japanese)
8. Kotler, P. and Lee, N.: Corporate social responsibility: Doing the most good for your company and your cause. John Wiley & Sons, Hoboken, NJ (2005).
9. Lawrence, A. and Weber, J.: Business and society: Stakeholders, ethics, public policy. 14th edn. McGraw Hill, New York (2013)
10. Murata, K., Adams, A. A., Orito, Y., Fukuta, Y. and Yamazaki, T.: Attitudes towards business information ethics in Japan. (in preparation)
11. Bridgman, T., Cummings, L. and McLaughlin, C.: Restating the case: How revisiting the development of the case method can help us think differently about the future of the business school. Academy of Management Learning & Education 15(4), 724–741 (2016)
12. Brey, P. and Dainow, B.: Ethics by design for artificial intelligence. AI and Ethics 4, 1265–1277 (2024)
13. Kaptein, M.: A Paradox of ethics: Why people in good organizations do bad things. Journal of Business Ethics, 184, 297–316 (2023)

ON THE CONSUMER EXPECTATIONS AND CONCERNS OF CONSUMERS REGARDING PERSONAL DATA USE AND COLLECTION INSIGHTS FROM WEB SURVEY RESULTS

HIROSHI KOGA[1]

This study will examine the existing state and issues based on the results of a questionnaire survey to ascertain consumer attitudes toward the use of user information such as action histories. In particular, the author would discuss the su
sumers' expectations and concerns about the use of personal data.

Background

Our interest is to understand the actual status of consumer awareness of the collection and use of consumers' personal information by companies and to examine the issues involved. In developing the questionnaire, we focused on the following three points.

1. Consumers' precautionary attitude toward personal information;
2. Desired content of control of personal information;
3. Degree of concern about the use of personal information by companies.

Methods

Based on the awareness of the above issues, we decided to conduct a questionnaire survey of general consumers. The survey was conducted using a self-administered questionnaire tool (Freeasy) provided by I-Bridge Corporation. A total of 1204 people, 86 male and 86 female, of all ages, were surveyed from among the monitors registered with the site. The mean age of the respondents was 45.41, S.D.=19.11. It should be noted, however, that the representativeness of the respondents is limited because of the survey method used.

The question that is the subject of analysis in this study is about "the degree of concern about the use of personal information by companies. Specifically, (1) expectations and concerns about the overall trend of companies' use of personal information, (2) attitudes toward the acquisition of personal information by companies, and (3) attitudes

1 FACULTY OF INFORMATICS, KANSAI UNIVERSITY, 3-3-35 YAMATE-CHO, SUITA-SHI, OSAKA 564-8680, JAPAN.

toward the use of personal information by companies were each asked in a four-point test. A multiple-answer system was employed, in which respondents were asked to select all applicable items from a list of 10 items, including “other.

Summary of survey result

Expectations and Concerns about Personal Data Use

Consumers were asked how they feel about companies’ use of their personal information. The results show that, overall, the most common response was “somewhat anxious” (33.1%), which together with the second most common choice (28.2%), “very anxious,” accounted for over 60% of the responses. This means that less than 40% of consumers have high expectations. Furthermore, the selection rate for “high expectations” was the smallest (12.3%). Overall, it can be said that consumers are not so optimistic and are feeling a little anxious, despite the importance of data science. If we look at the differences by gender, males had a higher selection rate for “expectation > anxiety” and females had a higher rate for “anxiety > expectation” (a chi-square test was conducted, and a significant difference was obtained, $p<0.001$). In Buddhism, the tendency to “look forward to the future” is called the masculine principle, while the tendency to value “immediate happiness” is called the feminine principle (and it is argued that it is important to possess both male and female principles). Coincidentally, the results of this survey coincide with this point.

Table 1. The relationship between “expectations vs. anxieties” and “gender differences”. By age group, the selection rate of “expectation > anxiety” was higher among teenagers (20.3%). On the other hand, those in their 50s (70.3%) and 70s and older (69.2%) selected “Anxiety > Expectation” more frequently (based on age, a chi-square test was conducted, and $p<0.001$ was obtained, indicating a significant difference).

	expectation > anxiety				Total
	n	%	n	%	n
10s	87	50.6%	85	49.4%	172
20s	74	43.0%	98	57.0%	172
30s	70	40.7%	102	59.3%	172
40s	66	38.4%	106	61.6%	172
50s	51	29.7%	121	70.3%	172
60s	64	37.2%	108	62.8%	172
over 70s	53	30.8%	119	69.2%	172
Total	465	38.6%	739	61.4%	1204

Table 2. Relationship between “Expectations vs. Anxiety” and “by age group”

Next, we turn our attention to the results of the cross tabulation with the frequency of smart-

phone use. 40.0% of those who selected "expectation > anxiety" responded that they "look at (smartphones) whenever I have free time. Conversely, for those who chose "Anxiety > Expectation," a higher percentage of respondents "rarely look at it" and "don't have a smartphone. A chi-square test was conducted, and a significant difference was obtained with $p<0.001$.

	Total	Male	Female
n	465	263	202
%	100.0%	56.6%	43.4%
n	739	339	400
%	100.0%	45.9%	54.1%
n	1204	602	602
%	100.0%	50.0%	50.0%

Relationship with the right to control one's own information

In Japan, the 2021 revision of the Personal Information Protection Law has brought renewed attention to the idea of understanding privacy as "the right to control information about oneself" (Itakura, 2021).

What kind of control do consumers want? In order to grasp the actual situation, we asked "desirable control contents of personal information. Furthermore, when we looked at the cross tabulation of the results and "expectation vs. anxiety," we found that those who chose "expectation > anxiety" showed a high selection rate for the "confirmation" and "correction" items. People who expect data utilization not only tolerate the use of their personal information but also demand the ability to review and correct the data held by companies when necessary. This aligns with the concept of the right to self-information control.

On the other hand, those who selected "Anxiety > Expectation" showed higher selection rates for "Erasure," "Disclosure to a third party," and "Limitation of retention period. The chi-square test revealed that "Confirmation" ($p<0.001$), "Correction" ($p<0.001$), "Erasure" ($p<0.001$), "Restriction on provision to third parties" ($p=0.002$) and "Restriction on retention period" ($p=0.010$), and significance was obtained for these five items. The importance of review and correction, similar to the expectations of those with high expectations, indicates that in the future use of personal information by companies, allowing consumers to review and correct their personal data will be an essential factor for businesses. Furthermore, it suggests that there is a tendency for consumers to want the use of their personal data to be restricted to the company that holds it, to avoid retaining it beyond the service period, and to request the right to erasure.

Attitudes toward Collection and Use of Personal Information

Next, we present the results of a survey on consumer attitudes toward the collection and use of personal information by companies, such as Web site browsing, search keyword history, and shopping records.

Specifically, the survey asked consumers about their opinions on (1) the collec-

tion activities themselves and (2) the use of such information for product recommendations.

Overall, the item with the highest selection rate regarding the collection of personal information by companies was “I wish they would stop if possible” (29.2%). This was followed by “I think it is inevitable,” and “I don’t understand the meaning of collecting personal information,” which together accounted for more than 70% of the total (25.6% and 15.6%, respectively).

Of the 30% of the total positive opinions, about 10% (10.5%) were strongly in favor of “I would be willing to provide it if it has merit,” and less than 20% (18.0%) selected “It is essential for me to receive appropriate services.

No significant difference between male and female respondents was found for these items ($p=0.168$).

On the other hand, a chi-square test for age showed a significant difference ($p<0.001$). Teenagers had more positive opinions, while those in their 20s were the most likely to deny the need to collect information, and those in their 50s and older were more likely to say “it can’t be helped” or “I wish they’d stop if possible.” It may be that people in their 20s have a strong aversion to the collection of personal information because this is a time when their lifestyles are changing dramatically.

Concerns about the collection and use of personal information

Next, we examine consumers’ concerns about the collection and use of personal information by companies.

The questionnaire asked consumers to select all applicable items from the following list regarding their “concerns about companies’ collection of personal information such as purchase histories and Web site browsing histories.

1. “Insufficient explanation” regarding the use of data
4. “Lack of veto power” of consumers
5. “Information management system information” of companies
6. Concerns about “resale” of information as a “source of profit
7. Fear of “anonymized” processing
8. Doubts about “collecting more information than necessary
9. Discomfort with “persistent use of old information
10. Business ethics (profit first)
11. Indication of “offensive advertisements
12. Fear of “profiling
13. Other (free answer)

Overall, the item with the highest selection rate was “management system”

(41.4%). This was followed by "Profiling" with the highest selection rate (34.6%). Items with selection rates exceeding 25% were "Insufficient explanation (28.8%)," "Profit source (resale)" (28.2%), and "Lack of veto power (25.3%). Other items were selected by around 20% of the respondents.

First, no gender differences were found in the selection rate of anxiety. As for age differences, statistical significance was found for individual identifiability ($p=0.002$) and lack of veto power ($p=0.005$).

Relationship between "expectation vs. anxiety

The relationship between "expectation vs. anxiety" and "expectation vs. anxiety" is further shown above. A chi-square test was conducted, and significant differences were obtained for the following six items. In order of overall selection rate, they are: "Profiling" ($p<0.001$), "Repeated use of old information" ($p=0.032$), "Anonymization technology" ($p=0.009$), "Collection of more information than necessary" ($p=0.001$), "Offensive advertisement" ($p=0.03$) and "Profit first" ($p<0.001$). (p-values in parentheses). Again, a one-way analysis of variance was performed and statistical significance was observed for similar items. For these items, the selection rate was higher among Kanto companies, which felt more anxious. To this extent, it can be concluded that anxiety regarding the use of personal information is caused by these factors.

Relationship with "Desirable control contents of one's own information

Finally, we show the relationship with the desirable control contents of self-reported information.

However, since the desired "control contents" can have up to three multiple responses and the "anxiety factors" can also have multiple responses, a chi-square test using a crosstabulation table is not appropriate. Therefore, in this paper, a chi-square test between a set of items was conducted by taking one item from each of the "control contents" and "anxiety factors. The overall trend revealed that the selection rate of "limitation of holding period" was low for seven of the control contents. In addition, the selection rate of "elimination" was high for all the anxiety factors, obtaining a significant difference. This result is appropriate considering the recent attention to the "right to be forgotten.

Conclusions and remarks

In this study, we have examined the results of a survey of consumer attitudes toward the collection and use of personal information. When asked which they felt more strongly about the data society, "expectations such as improved convenience" or "risks and concerns such as information leakage and invasion of privacy," men tended to select "expectations" and women "concerns. Those who selected "Expectation" used the four major social networking services in both real and anonymous mode, and wanted "Confirmation" and "Modification" in terms of self-information control. On the other hand, those who were "anxious" tended not to use SNSs, and they emphasized "deletion," "limitation of retention period," and "limitation of provision to a third party" as their right to control their own information. In addition, those who feel "uneasy" are more

likely to answer “hobby” or “none” as the contents of information that can be entrusted to “information banks.

The importance of data may increase, but it will not decrease in the future. Under these circumstances, we hope that the results of this survey will provide some insight into what factors contribute to expectations and what factors are effective in alleviating anxiety about the use of data.

Acknowledgements: This work was supported by JSPS KAKENHI Grant Number JP23K01620 and the Kansai University Fund for Domestic and Overseas Research Support Fund, 2023.

References

1. Itakura, Y. (2021) “The right to control one’s own information in an AI and big data society,” National Life (Web Edition: Reading, Understanding, and Thinking about Consumer Issues), Vol. 108, pp. 11-14. (in Japanese)
2. Ueki, M. (2004) “Views on men and women in Buddhism: The idea of gender equality from primitive Buddhism to the Lotus Sutra,” Iwanami Shoten (later renamed “Overcoming discrimination: The view of humanity in primitive Buddhism and the Lotus Sutra,” Kodansha, 2018). (in Japanese)
3. Koga, H. (2023) “Expectations and Concerns Regarding the Collection and Use of Personal Information: Considerations Based on the Results of a Web Survey,” Jornal of Informatics [Kansai University, Departmental Bulletin Paper], Vol.57, pp.61-80. (in Japanese)

CONSUMER PERCEPTIONS OF PERSONAL DATA TRUST BANKS IN JAPAN, RESULTS OF WEB-BASED QUESTIONNAIRE SURVEY

Hiroshi Koga[1]

Introduction

In recent years, data has been referred to as the new "oil" for businesses. Especially for companies targeting consumers, personal data such as purchase history has become an essential resource for executing efficient and effective marketing. In response to this, the Japanese government significantly revised the "Personal Information Protection Act" in 2022, aiming not only to ensure strict management of personal information but also to encourage its active use. As a result, the focus has shifted toward the utilization of personal information, while still maintaining the protection of individual rights and interests.

In light of these trends, information banks have garnered attention as a system that balances the protection and utilization of personal information.
The aim of this study is to clarify the current status and challenges surrounding the expectations and concerns related to the use of consumer personal data (such as behavioral history data) in Personal Data Trust Banks, based on the results of a web-based survey.

Significance of Information Banks

Platform companies, represented by GAFA(M), have increasingly been collecting vast amounts of personal information. Zuboff refers to this type of society as "surveillance capitalism." As it is self-evident, the aggregation of vast personal data such as purchasing history by specific platform companies is considered undesirable for the healthy development of capitalism. This is because monopolistic companies could become self-serving. Perhaps for this reason, regulations against platform companies are being considered in Japan.

On the other hand, the government is actively promoting information banks as a framework for managing and utilizing personal information securely. An information bank is a service that provides individuals and businesses with a system for managing and operating their data, and sharing it with others when necessary. It particularly focuses on the protection of personal data and privacy, aiming to allow individuals to control their own information.

The expected functions of information banks are as follows: Specifically, they are expected to undertake the following tasks upon commission by consumers: (1-1)

1 Faculty of Informatics, Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka 564-8680, Japan

managing and protecting personal data, (1-2) supporting the sharing and provision of data, (1-3) managing privacy, and (1-4) providing user rewards, such as points. By performing these functions, information banks are expected to offer the following benefits to consumers: (2-1) strengthening the control rights over personal information, (2-2) ensuring safety, (2-3) effective utilization of personal data, and (2-4) secondary effects of personal data utilization (e.g., point rewards). On the other hand, for businesses, the expected benefits include: (3-1) compliance with the Personal Information Protection Act, (3-2) establishing a foundation for personal data utilization, such as anonymization, and (3-3) ensuring transparency. In this way, by taking security measures while creating many benefits from personal information, information banks are expected to become a new industrial foundation for the future. As a result, individuals will be able to control their information, and safe data exchange with businesses will lead to enhanced privacy protection and the creation of new business opportunities.

However, from the perspective of personal information protection, to operate information banks, which are beneficial to consumers, businesses, and society as a whole, it is essential to obtain user consent. Users need to be transparently informed about how their data will be utilized and where it will be used. For this purpose, the design of the information bank system has been the subject of active discussions.

As preparation for this system design discussion, it is important to clarify what types of personal information consumers are willing to entrust to information banks. Therefore, the purpose of this paper is to conduct a survey on the awareness of Personal Data Trust Banks in Japan, present the results, and discuss them.

Methods

This study employed a questionnaire survey of general consumers. The survey was conducted using a self-administered questionnaire tool (Freeasy) provided by I-Bridge Corporation. A total of 1204 people, 86 male and 86 female and 86 male and female by age, were surveyed from among the monitors registered with the site. The survey period was from December 9, 2022, to December 10, 2022. The average age of the respondents was 45.41, with a standard deviation of 19.11.

However, it is important to note that because a web survey company was used, there are limitations to how representative the respondents are.

Summary of survay results

Overview of the Survey of PDTB

In the survey, respondents were asked to indicate the information they would be willing to entrust to a Personal Data Trust Bank from among the following items (multiple responses were acceptable).

1. “Personal attribute information” such as address and gender
2. “Family personal information,” such as family structure and occupation
3. “Financial information” such as income

4. “Personal identification codes” such as account or credit card numbers
5. “Shopping history information” (purchase history information)
6. “Movement information” mediated by smartphones
7. “Vital information” such as pulse, blood pressure, and body temperature
8. “Genetic information” such as medical history
9. “Hobbies” and other information
10. Others (free description)

Overall characteristics

Overall, the item with the highest selection rate is “Hobbies” (38.5%). This was followed by “personal attributes” (29.7%) and “purchase history” (29.3%). The item with the lowest selection rate was “Other” (5.9%: n=71). The respondents were asked to freely select “Other. The breakdown of the responses is as follows: “None,” 67 respondents chose “None,” 2 respondents chose “Don’t know,” and 2 respondents chose “Other. The following analysis is based on the 67 respondents who answered “none” in the free response section.

Table 1. The overall characteristics of the information items stored in the PDTB
The only significant gender difference was found in the “purchase history” category (p=0.004 by the chi-square test).

	n	%	Famale %
“Hobbies” and other information	464	38.5%	51.5%
“Personal attribute information” such as address and gender	357	29.7%	53.2%
“Shopping history information” (purchase history information)	353	29.3%	56.4%
“Vital information” such as pulse, blood pressure, and body temperature	221	18.4%	51.6%
“Movement information” mediated by smartphones	195	16.2%	44.6%
“Family personal information,” such as family structure and occupation	194	16.1%	52.6%
“Personal identification codes” such as account or credit card numbers	165	13.7%	50.9%
“Financial information” such as income	151	12.5%	47.0%
“Genetic information” such as medical history	143	11.9%	44.8%
Others (free description)	71	5.9%	52.1%

Table 2. The age-based distribution of the information items stored in the PDTB

	10s	20s	30s	40s	50s	60s	70s & over	Total
“Personal attribute information” such as address and gender	33.7%	20.9%	23.8%	27.3%	26.2%	33.1%	42.4%	29.7%
“Family personal information,” such as family structure and occupation	22.1%	14.0%	14.5%	15.1%	16.3%	14.5%	16.3%	16.1%
“Financial information” such as income	11.6%	15.7%	11.6%	14.5%	12.2%	11.0%	11.0%	12.5%
“Personal identification codes” such as account or credit card numbers	15.7%	14.5%	15.1%	11.0%	11.6%	14.0%	14.0%	13.7%
“Shopping history information” (purchase history information)	31.4%	23.8%	25.0%	27.3%	26.7%	26.7%	44.2%	29.3%
“Movement information” mediated by smartphones	23.3%	14.5%	13.4%	19.2%	11.0%	12.2%	19.8%	16.2%
“Vital information” such as pulse, blood pressure, and body temperature	18.6%	9.9%	12.8%	8.1%	10.5%	6.4%	16.9%	11.9%
“Genetic information” such as medical history	38.4%	39.5%	37.8%	33.1%	38.4%	37.8%	44.8%	38.5%
“Hobbies” and other information	1.7%	6.4%	7.0%	4.1%	11.6%	8.7%	1.7%	5.9%
Others (free description)	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Chi-square tests were conducted for each age group, and significant differences were obtained for "personal attributes" ($p<0.001$), "purchase history" ($p<0.001$), "vital information" ($p<0.001$), "medical history and genetic information" ($p=0.003$), and "none" ($p<0.001$) (p values in parentheses (p-values in parentheses).

Next, we discuss the results based on the crosstabulations by age group. Let us point out the characteristics of the results as follows. Personal attributes," which ranked second in the overall selection rate, was selected most frequently by respondents in their teens and those in their 60s or older. Next, the selection rates for "Purchasing history" and "Vital information" were high for respondents in their 70s, and slightly higher for respondents in their teens. The selection rate for "Medical history/previous diseases/genetic information" was high among teens, while the selection rate for "None" was high among those in their 50s. To simplify the results without fear of misunderstanding, more respondents in their teens and those in their 60s or older selected items that can be entrusted to the Personal Data Trust Bank, while those in their 50s selected fewer items.

Relationship with "Expectations vs.

Does the information that consumers are willing to entrust to a Personal Data Trust Bank have any relationship with their "expectations and fears" of the data society as a whole? To examine this question, we conducted a cross tabulation of "information that can be entrusted" and "expectation vs. anxiety.

A chi-square test was conducted, and significant differences were obtained for nine items (including "other") except for "medical history/previous illness/genetic information" ($p=0.015$).

Those who selected "expectation > anxiety" selected the following seven items more frequently. The items selected most frequently by those who selected "Expectation > Anxiety" were "Purchase history," "Personal attributes," "Vital information," "Transportation information," "Family information," "Account information," and "Income/assets information," in descending order. Those who selected "expectation > anxiety" selected an average of 2.1 items of "information that can be entrusted to the bank. On the other hand, those who selected "Anxiety > Expectation" selected an average of only 1.8 items.

In addition, those who selected "Anxiety > Expectation" selected "Hobbies" and "None" at a higher rate than those who selected "Expectation > Anxiety," a result that is symmetrical with those who selected "Anxiety > Expectation. The data also revealed that those who were anxious were reluctant to use the Personal Data Trust Bank.

Relationship with "Desirable Control Contents

Finally, we examine the relationship with "desirable control contents" (Table 16). First, "hobbies" was significantly different from "deletion," "collection restrictions," and "retention period restrictions. All these items were selected more frequently by the respondents with "anxiety > expectation". Therefore, it can be inferred that those who were anxious about the collection and use of personal information selected "Hobbies" as the safest or least sensitive item when using the Personal Data Trust Bank.

Similarly, in the case of "None," which was selected by a high percentage of respondents with "Anxiety > Expectation," there was a significant difference between "Deletion" and "Collection Restrictions. However, this item was not included in the questionnaire and was extracted from the free descriptions in the "Others" section. Therefore, it is safe to assume that these are the responses of those who are most reluctant to use the Personal Data Trust Bank. If so, we can understand that "deletion" and "collection restriction" are the most important control items for those who are reluctant (dare we say negative) to the idea of a Personal Data Trust Bank.

Next, let us consider the "information that can be entrusted to the bank," which was selected most frequently by those who selected "expectation > anxiety.
The relationship with the "limitation of retention period," which was selected by a high percentage of respondents with "expectation > uneasiness" was examined, and significant differences were obtained for the following five items. The significant differences were found for the following five items: "personal information," "purchase history," "family information," "account information," and "income/assets. However, the selection rate was low for all items.

Comparing each item with the "desirable control contents," for which significant differences were obtained, we can see the characteristics of personal information that can be entrusted to the Personal Data Trust Bank. For example, for "personal information," the respondents want to be able to limit the content of information collected and to be able to disclose and correct the content of that information. In the case of "purchase history," the respondents want not only collection but also restriction of use, and not only correction but also suspension (deletion) of use. For "family information," the respondents want not only the collection but also the restriction of use and the suspension of use (deletion) in addition to the requirements for "personal information. In the case of "account information" and "income/assets," only "correction" seems to be emphasized.

Furthermore, Table 16 is also suggestive for the four items for which no significant difference was obtained for "limitation of holding period. We have already mentioned "None". Vital information" was the only item among the 'information contents' related to the Personal Data Trust Bank that showed a significant difference from 'Restrictions on provision to third parties. Not only were the selection rates of "Restrictions on collection" and "Confirmation and correction" of vital information high, but also the respondents' desire not to have their information used for health-related advertisements and sales can be seen.

The "movement information" was significantly different from the three items of "personal information" except for "limitation of retention period" from "control contents" which was significantly different from "personal information.
Finally, for "medical history, pre-existing conditions, and genetic information," the only "desirable control item" that showed a significant difference was "limitation of collection.

Conclusions and Remarks

In the above, we have examined the results of a questionnaire survey on consumer attitudes toward PDTB, which have been the focus of much attention in recent

years. In the future, we would like to investigate the relationship with consumers' security concepts such as cyber hygiene.

Acknowledgements: This work was supported by JSPS KAKENHI Grant Number JP23K01620 and the Kansai University Fund for Domestic and Overseas Research Support Fund, 2023.

References

1. Itakura, Y. (2021) "The right to control one's own information in an AI and big data society," National Life (Web Edition: Reading, Understanding, and Thinking about Consumer Issues), Vol. 108, pp. 11-14. (in Japanese)
2. Sato, Y and Sunada, Y (2020) "How to use the Personal Data Trust Bank – Overview of the new system to handle personal data," The Journal of Information Science and Technology Association, Vol.70, No.11, pp.546-551 (in Japanese)
3. Koga, H. (2023) "Expectations and Concerns Regarding the Collection and Use of Personal Information: Considerations Based on the Results of a Web Survey," Jornal of Informatics [Kansai University, Departmental Bulletin Paper], Vol.57, pp.61-80. (in Japanese)

BEYOND DELETION: THE AFTERLIFE OF IMAGES IN ALGORITHMIC GOVERNANCE

Klara Källström[1], Marie Eneman[2], Jan Ljungberg[3]

Introduction

Contemporary forms of algorithmic governance increasingly challenge established assumptions about autonomy, particularly in relation to how personal data and 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 for human interpretation, are now processed, fragmented, and recomposed as data points within computational infrastructures of classification and control (Amoore, 2020). This ontological shift marks the industrialization of images, whereby they are no longer encountered as visual artefacts but operate as functional components in automated decision-making systems (Farocki, 2004). In such environments, images and personal data are rendered equivalent—both reduced to informational units circulating through networks. This abstraction has significant implications for understandings of privacy, as the dispersal of visual data across algorithmic infrastructures renders it opaque, untraceable, and resistant to withdrawal (Dewdney & Sluis, 2023).

This paper investigates ontological and epistemological shifts through the case of the Swedish police's use of Clearview AI's facial recognition technology. It examines how images, once absorbed into algorithmic systems, are reconstituted as dispersed computational entities, challenging the legal and technical viability of regulatory mechanisms such as the Right to Be Forgotten (Kosta, 2022). When images are embedded within predictive and classificatory infrastructures, their withdrawal reveals the always-already unstable nature of archival systems—where erasure is not only technically complex but historically contingent. This raises the research question: How does the industrialisation of images shape the ability to uphold the Right to Be Forgotten in archival practices?

Through an analysis of official statements and administrative records, the study interrogates how visual data persists within algorithmic governance and considers what it means to "forget" in a system where the image is not a visible artefact but a distributed signal structured by data flows and institutional power (Bowker, 2005; Dewdney & Sluis, 2023; Jasanoff, 2004). In this context, archival practices must be thought of differently—not as systems for storing and retrieving static records but as dynamic infrastruc-

1 Department of Applied Information Technology, University of Gothenburg
2 Department of Applied Information Technology, University of Gothenburg
3 Department of Applied Information Technology, University of Gothenburg

tures entangled with algorithmic processes of extraction, recomposition, and circulation. As the archive becomes operational rather than representational, new frameworks are needed to address how memory, visibility, and erasure function under computational governance.

Theoretical foundation

Informational privacy is often framed as the individual's capacity to dissociate from persistent digital traces (Floridi, 2015; Richards, 2022; Véliz, 2020). This framing assumes a coherent, autonomous subject capable of controlling personal data. Gilles Deleuze's concept of the dividual challenges this, proposing that identity in computational systems is fragmented and operationalized through data points, behaviours, and metrics (Deleuze, 1992). Within this framework, privacy concerns shift from ownership of discrete data to the aggregation and circulation of informational fragments used for classification and prediction (Bowker & Star, 1999; Amoore, 2024; Lyon, 2003). A similar ontological shift characterises the treatment of images: once processed algorithmically, images are transformed into dispersed data points embedded within computational infrastructures (Dewdney & Sluis, 2023; Paglen, 2016). Like personal data, the image becomes fragmented, recontextualised, and ultimately inaccessible as a coherent object.

As algorithmic systems increasingly shape how data is processed and operationalised within governance frameworks, the inherent instability of autonomy—central to rights such as the Right to Be Forgotten—comes to the fore (Ananny & Crawford, 2018; Pasquale, 2015). While the GDPR's Right to Be Forgotten seeks to mitigate the persistence of digital traces, it rests on the assumption that deletion restores privacy. This falters in distributed and predictive infrastructures where data—visual or otherwise—is continuously replicated, recombined, and embedded in classificatory systems. Deletion becomes not a definitive act but a contingent intervention shaped by archival logics and infrastructural resistance (Bowker, 2005; Kosta, 2022; Mantelero, 2013).

These challenges demand critical engagement with algorithmic infrastructures that govern behaviour through data flows (Jasanoff, 2004; Latour, 2005), emerging from a genealogy of classificatory and surveillance practices shaped by historically contingent regimes of power and knowledge (Foucault, 1977). Archival systems have long functioned as instruments of institutional power, shaping access, memory, and control (Bowker, 2005; Geoghegan, 2023), with photography playing a central role in biometric surveillance and criminological classification (Lyon, 2014; Sekula, 1986)—logics that persist in contemporary facial recognition technologies (Ananny & Crawford, 2018; Crawford, 2021). The shift from physical archives to digital repositories—and more recently to algorithmically driven infrastructures—has intensified these dynamics: unbound by spatial constraints, algorithmic systems enable indefinite data retention and circulation, complicating regulation around deletion, access, and classification (Bowker & Star, 1999; Paglen, 2016). Regulatory strategies must therefore confront the entanglement of algorithmic infrastructures with archival practices, surveillance, and the politics of data retention.

Research Design

Setting

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 (Rezende, 2020; Shepherd, 2024). Images become functionally opaque, mobilised not as representations but as operational inputs in automated recognition systems. The controversy intensified in 2020 when BuzzFeed News leaked Clearview's client list, revealing use by law enforcement agencies in the U.S., Canada, and several European countries, including Sweden. In Canada, Finland, and Sweden, officers accessed the system without institutional approval, blurring boundaries between state oversight and commercial infrastructures (Amoore, 2020; Eneman et al., 2022; Shepherd, 2024).

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 highlights the difficulty of enforcing protections like the Right to Be Forgotten when images are embedded in distributed systems that resist deletion (Mantelero, 2013). IMY ordered affected individuals to be notified and transferred data erased (Eneman et al., 2022; Kosta, 2022).

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)

Document Collection and Analysis

This study draws on public records from the IMY, the Police Authority, and the Administrative Court in Stockholm, obtained under Sweden's Principle of Public Access to Information, grounded in the Freedom of the Press Act (1949:105). While this principle

ensures transparency, it is increasingly strained by digital infrastructures where traceability and accountability are diminished (Ananny & Crawford, 2018; Bowker, 2005).

The analyzed documents (Bowen, 2009) reveal how facial recognition technologies intersect with legal reasoning and classification regimes. They show how images, once housed in institutional archives, now circulate through algorithmic systems that challenge conventional understandings of regulation, erasure, and control (Eneman et al., 2022).

Concluding Remarks

This analysis 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 industrialisation. 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 (Amoore, 2020).

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—no longer stored but used, and deeply entangled with material, geopolitical, and logistical networks of appropriation and circulation (Crawford & Joler, 2021). In this context, the image becomes part of a global supply chain of data operations to be mined, reformulated, and propagated. In this Farockian sense, archives cease to function as static repositories and instead become industrialised systems of capture and recomposition, where images are not preserved but enacted upon (Amoore, 2024; Bowker, 2005; Farocki, 2004).

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) underscore this disjunction. The directive to delete and notify presumes a subject who can be located, addressed, and restored to privacy—assumptions incompatible with algorithmic infrastructures where data persists not as content, but as action.

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.

References

1. Amoore, L. (2020). Cloud ethics: Algorithms and the attributes of ourselves and others. Duke University Press.

2. Amoore, L., Campolo, A., Jacobsen, B., & Rella, L. (2024). A world model: On the political logics of generative AI. Political Geography, 113, 103134. https://doi.org/10.1016/j.polgeo.2024.103134
3. Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645
4. Bowen, G. A. (2009). Document analysis as a qualitative research method. Qualitative Research Journal, 9(2), 27–40. https://doi.org/10.3316/QRJ0902027
5. Bowker, G. C. (2005). Memory practices in the sciences. MIT Press.
6. Bowker, G. C., & Star, S. L. (1999). Sorting things out: Classification and its consequences. MIT Press.
7. Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press. https://doi.org/10.12987/9780300252392
8. Crawford, K., & Joler, V. (2021). Anatomy of an AI system: The Amazon Echo as an anatomical map of human labor, data, and planetary resources. In K. Crawford, Atlas of AI: Power, politics, and the planetary costs of artificial intelligence (pp. 197–217). Yale University Press.
9. Deleuze, G. (1992). Postscript on the societies of control. October, 59, 3–7.
10. Dewdney, A., & Sluis, K. (Eds.). (2023). The networked image in post-digital culture. Routledge.
11. Eneman, M., Ljungberg, J., Raviola, E., & Rolandsson, B. (2022). The sensitive nature of facial recognition: Tensions between the Swedish police and regulatory authorities. Information Polity, 27(2), 219–232. https://doi.org/10.3233/IP-211538
12. Farocki, H. (2004). Phantom images. Public, 29, 12–22.
13. Farocki, H. (2001–2003). Eye/Machine I–III [Video essay series]. Harun Farocki Filmproduktion
14. Floridi, L. (2015). The onlife manifesto: Being human in a hyperconnected era. Springer. https://doi.org/10.1007/978-3-319-04093-6
15. Foucault, M. (1977). Discipline and punish: The birth of the prison (A. Sheridan, Trans.). Pantheon Books.
16. Geoghegan, B. D. (2023). Code: From information theory to French theory. Duke University Press.
17. Gitelman, L. (2013). Raw data is an oxymoron. MIT Press.
18. Jasanoff, S. (2004). States of knowledge: The co-production of science and the social order. Routledge. https://doi.org/10.4324/9780203413845
19. Kosta, E. (2022). Algorithmic state surveillance: Challenging the notion of agency in human rights. Regulation & Governance, 16(1), 212–224. https://doi.org/10.1111/rego.12331
20. Latour, B. (2005). Reassembling the social: An introduction to actor-network theory. Oxford University Press.
21. Lyon, D. (2014). Surveillance studies: An overview. Polity Press.
22. Mantelero, A. (2013). The EU proposal for a General Data Protection Regulation and the roots of the 'Right to Be Forgotten.' Computer Law & Security Review, 29(3), 229–235. https://doi.org/10.1016/j.clsr.2013.03.010
23. Paglen, T. (2016). Invisible images: Your pictures are looking at you. The New Inquiry. Retrieved March 10, 2025, from https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/
24. Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061
25. Rezende, I. N. (2020). Facial recognition in police hands: Assessing the 'Clearview case' from a European perspective. New Journal of European Criminal Law, 11(3), 375–389. https://doi.org/10.1177/2032284420948161
26. Richards, N. M. (2022). Why privacy matters. Oxford University Press.

27. Rouvroy, A., & Berns, T. (2013). Algorithmic governmentality and prospects of emancipation: Disparateness as a precondition for individuation. (L. Carey-Libbrecht, Trans.). Réseaux, 177(1), 163–196. https://shs.cairn.info/journal-reseaux-2013-1-page-163?lang=en
28. Sekula, A. (1986). The body and the archive. October, 39, 3–64. https://doi.org/10.2307/778312
29. Shepherd, T. (2024). The Canadian Clearview AI investigation as a call for digital policy literacy. Surveillance & Society, 22(2), 179–191. https://doi.org/10.24908/ss.v22i2.16300
30. Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research, 7(1), 111–134. https://doi.org/10.1287/isre.7.1.111
31. Véliz, C. (2020). Privacy is power: Why and how you should take back control of your data. Bantam Press.

A CASE AGAINST THE FEASIBILITY OF AI CONSCIOUSNESS (AIC)

Bo Kampmann Walther[1] [0000-0003-0600-1779]

Anne Gerdes[2] [10000-0002-2991-5074]

Keywords: Consciousness, Large Language Models, Embodiment.

Extended Abstract

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 (Chalmers, 2023). 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" (Chalmers, 2023). 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: to be conscious 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. Moreover, 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 against AI Consciousness (AIC) can still draw from objections to Artificial General Intelligence (AGI). We argue that consciousness, i.e., the ability to have subjective experiences, requires embodiment (Dreyfus et al., 1986) and embeddedness in social life (Collins, 2008). Our position does not exclude the possibility of animal consciousness, which is well-documented. For instance, "distant relatives of humans, Eurasian jays, show sensitivity to others' desires, even when they conflict with

1 University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark
2 University of Southern Denmark, Universitetsparken 1, 6000 Kolding, Denmark, gerdes@sdu.dk

the subject's own. Our closest relatives, the great apes, show sensitivity to perspectives that differ from their own, responding appropriately to others' false beliefs" (Krupenye and Call, 2019).

Against this backdrop, we outline classical arguments for and against AIC, ultimately concluding that AIC is not feasible.

Pro et con AIC

In his influential article, What is It Like to Be a Bat? (Nagel, 1974), 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 (Jackson, 1986). 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.

Chalmers (1886, 2008) 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. However, the "hard problem" concerns 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 concerns whether consciousness can emerge from silicon. Physicalists 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 (1993) 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?

Despite such arguments, we believe synthetic brains would function like stochastic parrots (Bender et al., 2021), lacking appropriate world models. Extending the argument about the infeasibility of AGI to consciousness, we draw on Dreyfus et al. (1986), who argue that AGI is not feasible because computes are disembodied. Collins (2008) adds that socialization, rather than embodiment, is the primary hurdle to AGI. A disabled person, for instance, can understand practices without directly experiencing them bodily. Consequently, being "embedded in the language of the society" and social life is what matters (Collins, 2008). Collins agrees, however, that AGI would need

a human-like body to join our practices and "share our concepts" (Collins, 2019). While Collins and Dreyfus differ on which aspects are most critical, they agree that intelligence necessitates both embodiment and embedding.

We extend this conclusion to consciousness: AIC cannot emerge in silicon or any other synthetic substrate. An AI lacks the ability to experience even the simple humor of the fruitarian date in the introduction. It is simply not here.

Moreover, there is a critical distinction between conceptualizing consciousness as a 'stacked' layering of intelligences, where consciousness emerges at the outcome of increasingly complex cognitive structures, and our position, which views consciousness as a transcendental precondition for subjective experience. In the former view, consciousness is treated as the culmination of intelligence, which could be referred to as the "Nirvana" of AI. However, we argue that consciousness is not an emergent end product but rather the foundational basis that enables any form of subjective or sentient-invariant experience. The various pros et cons are summarized in figure one below.

In line with Putnam (1964), we thus contend that the possibility of AIC calls for a decision rather than a discovery.

Table 1. Pro et con arguments regarding the feasibility of AI consciousness (AIC).

ARGUMENT TYPE	ARGUMENTS FOR AIC	ARGUMENTS AGAINST AIC
PHILOSOPHICAL	**Physicalism and Biophysical Processes**: Mental states are equivalent to biophysical processes that can, in principle, be represented in machine-readable forms, making AIC possible (Copeland, 1993).	**Nagel's Qualia Problem**: Subjective experiences (qualia) cannot be fully understood or reproduced, even with complete knowledge of neurophysiology (Nagel, 1974).
	Emergent Properties: If non-physical properties can arise from natural brains, they might also emerge from synthetic brains (Copeland, 1993).	**Jackson's Mary Thought Experiment**: Knowing all physical facts about a phenomenon does not equate to experiencing it (Jackson, 1986).
SCIENTIFIC CHALLENGES	**Hard Problems of Consciousness**: Though challenging, scientific advancements may one day address the subjective experience problem (Chalmers, 2023).	**Chalmers' Hard Problem**: The subjective nature of experience—why processing feels like something—is unlikely to have a scientific answer (Chalmers, 1996).
EMBODIMENT AND SOCIALIZATION	**Language and Intelligence Embedding**: Embedding in social languages and practices could hypothetically allow synthetic entities to develop human-like concepts (Collins, 2008, 2019).	**Embodiment**: Consciousness fundamentally requires a body to interact with the world (Dreyfus et al., 1986).
		Social Embeddedness: Intelligence and consciousness emerge from social interaction, not just embodiment, and AI lacks the capacity for social participation in human practices (Collins, 2008).
NATURE OF CONSCIOUSNESS	**Stacked Layers Model**: Consciousness could emerge as the cumulative result of increasingly complex cognitive structures.	**Transcendental Precondition**: Consciousness is foundational and enables subjective experience; it cannot be treated as an emergent phenomenon or end product.
COGNITIVE FUNCTIONALITY	**Machine Simulation**: Machines could simulate cognitive and emotional states sufficiently to mimic consciousness.	**Stochastic Parrots Argument**: AI, like stochastic parrots, lacks real-world understanding or experience necessary for consciousness (Bender et al., 2021).
EPISTEMOLOGICAL	**Imaginative Limitations**: The inability to imagine synthetic consciousness does not inherently rule out its feasibility (Copeland, 1993).	**Decision, Not Discovery**: The possibility of AIC is a philosophical decision rather than a scientific discovery, as consciousness transcends technical frameworks (Putnam, 1964).

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

References

1. 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)
2. Chalmers, D. J.: Could a large language model be conscious? arXiv preprint arXiv:2303.07103 (2023)
3. Chalmers, D., J.: The Conscious Mind – In Search of a Fundamental Theory. Oxford University Press, New York (1996)
4. Chalmers, D., J.: The Hard Problem of Consciousness. In: Velmans, M., Schneider, S. (eds.) The Blackwell Companion to Consciousness. Whiley-Blackwell, Boston (2008)
5. Collins, H. M.: Remembering Bert Dreyfus. AI & society, vol. 34, 373-376 (2019)
6. Collins, H. M.: Response to Selinger on Dreyfus. Phenom Cognitive Science, vol 7, 309-311 (2008)
7. Copeland, J.: Artificial Intelligence. Blackwell Publishing, Oxford (1993)
8. 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).
9. Jackson, F.: What Mary Didn't Know. The Journal of Philosophy, 83(5), 291–295 (1986)
10. Krupenye, R., Call, J.: Theory of mind in animals: Current and future directions. Wiley Interdisciplinary Reviews, Cognitive Science, 10(6), e1503-n7a (2019)
11. Putman, H., & Putnam, H.: Robots: Machines or Artificially Created Life? The Journal of Philosophy, 61(21), 668–691 (1964)

USER ENGAGEMENT AND BARRIERS IN THE STANDARDIZATION PROCESSES OF DIGITAL ID ARCHITECTURES

YOSHIAKI FUKAMI[1] [0000-0002-7838-8215]
KAZUE SAKO[2] [0000-0003-0890-8083]

Keywords: Digital Identity, ID architecture, standardization, data governance, multi-stakeholder.

Extended Abstract

Introduction

The increasing reliance on digital identity systems (DIS) across both public and private sectors has significantly transformed how individuals interact with digital services. Digital ID solutions aim to enhance security, protect user privacy, streamline service delivery, and enable seamless access to public and private services. However, the standardization processes behind these systems remain fragmented, with limited user engagement and insufficient collaboration among diverse stakeholders, including technologists, policymakers, and legal experts [1]

This fragmented approach has hindered efforts to create cohesive, user-centered digital identity frameworks. Regulatory interventions, such as regulatory sandboxes, have been proposed to address the challenges of governing data-driven innovations in smart cities [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 generative AI chatbots. 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. This challenge extends to digital identity systems, particularly in cases where automated decision-making and self-sovereign identity (SSI) models are deployed without clear governance structures.

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, JAPANS, KAZUESAKO@AONI.WASEDA.JP

Public and Private Sector Approaches to Digital Identity

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].

One critical aspect of the EUDIW is the implementation of Selective Disclosure technology, which allows users to reveal only specific attributes to service providers. This concept is central to privacy protection as it reduces the amount of personal information shared in transactions. However, a major concern with the EUDIW is that it does not ensure unlinkability, meaning that a user's interactions can be linked across different services. This raises significant privacy concerns, as linkable credentials could lead to user tracking and data aggregation by service providers.

Achieving these goals requires the development, implementation, and adoption of new technologies, many of which are developed and standardized by private-sector consortia. Organizations such as the Internet Engineering Task Force (IETF), the World Wide Web Consortium (W3C), and the OpenID Foundation have been actively working on the standardization of technologies related to identity management, including OAuth, Decentralized Identifiers (DIDs), and Verifiable Credentials (VCs). The IETF has developed protocols like OAuth 2.0, which is widely used for secure authorization. The W3C has been working on the DID and VC specifications to enable self-sovereign identity (SSI) models, while the OpenID Foundation has been focusing on protocols for federated identity and authentication.

However, these standardization efforts often proceed in parallel across multiple organizations, leading to a fragmented landscape. Policymakers face challenges in engaging with these processes due to the decentralized nature of discussions and the technical complexity of the topics involved. This fragmentation complicates the adoption of cohesive governance frameworks and highlights the importance of forums like the Internet Identity Workshop (IIW), which provides an open space for stakeholders to discuss and coordinate efforts across different standardization bodies.

Blockchain and Decentralized Identity: Ethical Considerations

The push for decentralized identity systems has often been driven by Web3 proponents, including blockchain advocates and cryptocurrency-related organizations. These groups emphasize the importance of user autonomy, data control, and decentralized trust models. However, Verifiable Credential (VC) technologies can be implemented without relying on blockchain, leading to debates about whether blockchain is necessary for decentralized identity systems.

Some stakeholders argue that using blockchain for digital identity systems poses risks related to privacy, scalability, and environmental sustainability [8]. For instance,

the immutability of blockchain records can conflict with the right to be forgotten under GDPR [9]. This issue raises questions about how personal data should be managed in decentralized systems, particularly in light of privacy regulations [10].

Moreover, there are concerns about the ethical implications of promoting blockchain-based identity systems, particularly when alternative, blockchain-free implementations exist. Critics argue that the focus on blockchain could divert attention from more privacy-preserving technologies and create technological lock-in, limiting future innovation.

The Role of the Internet Identity Workshop (IIW)

The Internet Identity Workshop (IIW) has emerged as a key forum for discussing digital identity standards and technologies. Given the fragmented landscape of identity-related standardization efforts, the IIW serves an essential function by bringing together stakeholders from different organizations to discuss emerging trends, technological innovations, and policy considerations.

Unlike formal standardization bodies, the IIW is a grassroots movement that emphasizes open participation and collaborative problem-solving. It provides a platform for technical experts, policymakers, and business leaders to engage in discussions on identity management systems, particularly in the context of self-sovereign identity (SSI) and decentralized identity architectures.

One of the critical challenges faced by policymakers is keeping pace with the fast-evolving technical standards. Many of these standards are developed through informal consortia and workshops rather than traditional, formal standardization processes. As a result, policymakers often struggle to understand and adopt emerging technologies in a timely manner. The IIW helps address this gap by providing regular updates on the latest developments and facilitating cross-sector collaboration.

However, the IIW's informal structure also presents challenges. Participation is largely limited to those who can attend in person, which can exclude stakeholders from developing countries and public-sector organizations. Additionally, the lack of formal governance raises questions about accountability and the legitimacy of decisions made during the workshops.

Despite these limitations, the IIW plays a crucial role in bridging the gap between technical experts and policymakers, ensuring that user-centered and privacy-preserving solutions are considered in the development of digital identity systems.

Conclusion

Digital Future research should focus on bridging the gap between private-sector innovation and public-sector governance, ensuring that digital identity systems are developed in a transparent, inclusive, and ethically responsible manner. By addressing the barriers to user engagement and fostering collaborative policymaking, stakeholders can work toward more harmonized governance frameworks that prioritize user autonomy, privacy protection, and data security.

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. The Review of Socionetwork Strategies, 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, pp. 180–202. Association of Pacific Rim Universities Limited (APRU) and Keio University (2020)
3. UNESCO: Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137.locale=en (accessed January 13, 2025)
4. OECD: Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449 (accessed January 13, 2025)
5. IEEE Standards Association: Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf (accessed January 13, 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. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed January 13, 2025)
8. Linn, L.A., Koo, M.B.: Blockchain for health data and its potential use in health IT and health-care-related research. In: ONC/NIST Use of Blockchain for Healthcare and Research Workshop, pp. 1–10. ONC/NIST, Gaithersburg, MD, USA (2016)
9. Finck, M.: Blockchain and the General Data Protection Regulation: Can distributed ledgers be squared with European data protection law? European Parliamentary Research Service (EPRS). https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634445/EPRS_STU(2019)634445_EN.pdf (accessed January 13, 2025)
10. 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)

ETHICAL ISSUES IN THE USE OF GENERATIVE AI CHATBOTS FOR THERAPEUTIC PURPOSES

ANNE GERDES[1] [10000-0002-2991-5074]

Keywords: Therapy Chatbots, Generative AI, Authenticity

Introduction

This paper starts by giving an overview of the pros and cons associated with the overall digital transformation of mental healthcare, particularly focusing on the therapeutic domain (sec. 2). In this context, the paper presents AI therapy bots and generative AI companion bots and explores their potential benefits and risks, including attention to relevant legislation (sec. 3). Next, the paper discusses the challenges of authenticity and empathy associated with the use of generative AI chatbots (sec. 4). The paper concludes (sec. 5) that the authentic therapist-client relationship is central to therapy. However, generative AI chatbots lack lived experience, remain unreliable, and risk offering harmful advice. Therefore, they are not suitable for therapeutic purposes.

The Digital Transformation of the Therapeutic Field

Mental health issues are growing significantly worldwide (WHO, 2022), 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 (Center for Digital Psykiatri). AI-based online therapy, possibly with psychological support, can reduce waiting times and alleviate mild forms of depression and anxiety (Hedman-Lagerlöf et al., 2023). 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.

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

1 UNIVERSITY OF SOUTHERN DENMARK, KOLDING 6000, DENMARK

apps appear more serious and professional than general apps, which can mislead users into believing they offer a confidential therapeutic space, even when their data is sold for other purposes (Martinez-Martin et al., 2018).

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.

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 (Lawrence et al., 2024) and assist with fighting alcohol abuse (Prochaska et al., 2021). Users already perceive them as empathetic. Likewise, 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 AI companions like Replika, 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 (Duffy, 2025). Google has disclaimed responsibility, noting that they are unrelated to Character.AI.

The unregulated market for low-quality mental health apps and chatbots will likely expand as developing such services becomes more effortless. However, EU legislation aims to mitigate harmful products by requiring large online platforms to comply with the Digital Service Act (DSA), protect personal data under GDPR, and follow a risk-based approach under the AI Act. Consequently, AI systems in healthcare are classified as high-risk, requiring proactive management of ethical and safety issues. In contrast, AI companions, like Replika, are considered low-risk systems with requirements such as informing users that they are interacting with a machine. Moreover, providers of such systems shall avoid exploiting vulnerable groups (e.g., children and mentally ill persons). Additionally, AI companions might also fall under the scope of the DSA if they are integrated into very large online platforms (VLOP). EU legislation can presumably promote the responsible development of 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.

Generative AI Chatbots – the Challenge of Authenticity and Empathy

Let us revisit the 1960s when computer scientist Joseph Weizenbaum developed the ELIZA program (Weizenbaum, 1988). This early chatbot included a version called DOCTOR, based on the principles of Carl Roers' psychotherapy. Client-centered therapy is relatively straightforward to simulate because the therapist repeats the client's own words.

Weizenbaum was taken aback by the attention ELIZA earned. In fact, he experienced three shocks: first, psychiatrists expressed admiration for the potential of automated therapy, with some even claiming that computers could eventually replace and surpass therapists. 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 (Weizenbaum, 1976).

The notion of human-centeredness is critical in the intersection of authenticity (Taylor, 1991), empathy, and generative AI chatbots. Some studies tentatively suggest that patients find language models more empathetic than interactions with doctors (e.g., Tu et al., 2024). 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" (Bender et al., 2021). Moreover, Turkle (2011) warns of an emergent "robotic moment", where we might prefer risk-free interactions with AI companions over demanding interpersonal relationships.

Conclusion

With their realistic communication capabilities, generative AI chatbots invite us to share our inner life with a stochastic parrot (Bender et al., 2021). However, the authentic connection between therapist and client remains a cornerstone of therapy, and generative AI chatbots lack the lived therapeutic experience. Moreover, they are unreliable and can provide harmful therapeutic advice. There is still a long way to go before language models can be responsibly used in therapeutic contexts.

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, last accessed 2025/03/17

2. 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)
3. Center for Digital Psykiatri https://psykiatrienisyddanmark.dk/afdelinger/specialisere-de-afdelinger-og-centre/center-for-digital-psykiatri, last accessed 2025/03/17
4. Chaturvedi, R., Verma, S., Das, R., & Dwivedi, Y. K.: Social companionship with artificial intelligence: Recent trends and future avenues. Technological Forecasting and Social Change, 193, 122634 (2023)
5. Duffy, C. CNN: 'There are no guardrails.' This mom believes an AI chatbot is responsible for her son's suicide. https://edition.cnn.com/2024/10/30/tech/teen-suicide-character-ai-lawsuit/index.html accessed 2025/04/17 (03.04. 2025).
6. 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 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065, last accessed 2025/03/17
7. 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) https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng, last accessed 2025/03/17
8. 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)
9. 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)
10. Prochaska, J. J., Vogel, E. A., Chieng, A., Kendra, M., Baiocchi, M., Pajarito, S., & Robinson, A.: A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study [Original Paper]. J Med Internet Res, 23(3): e24850 (2021)
11. Taylor, C.: The Ethics of Authenticity. Harvard University Press, Massachusetts (1991)
12. Tu, T., Palepu, A., Schaekermann, M., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Tomasev, N., Azizi, S., Singhal, K., Cheng, Y., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S. S., Semturs, C., Gottweis, J., . . . Natarajan, V.: Towards Conversational Diagnostic AI, http://arxiv.org/abs/2401.05654, (2024)
13. Turkle, S.: Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books, NY (2011)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. et al.: Attention is all you need. 10.48550/arXiv.1706.03762 (2017)
15. Weizenbaum, J.: ELIZA—a computer program for the study of natural language communication between man and machine, Commun. ACM, 9(1), 36–45 (1966)
16. Weizenbaum, J.: Computer power and human reason: From judgment to calculation. Freeman San Francisco (1976)
17. WHO: World mental health report: transforming mental health for all. World Health Organization. Jun 16, 2022. URL: https://www.who.int/publications/i/item/9789240049338, last accessed 2025/03/17

SUSTAINABLE FASHION TRENDS ON SECOND-HAND SHOPPING: THE ROLE OF SOCIAL MEDIA

Lídia Monteiro[1]
Orlando Lima Rua[2] [0000-0002-1593-7440]

Keywords: Sustainable Fashion, Second-Hand Shopping, Social Media, Trends, University Students.

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.

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

1Porto 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 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; 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; Research Center of Business Sciences (NECE), University of Beira Interior (UBI), 6200-209 Covilhã, Portugal, lncs@springer.com

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].

Literature Review

[4] 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 [7]. 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 [7]. 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 form 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?

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 [8]. 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. 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)

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.

Conclusion

This study provides a nuanced exploration of the interplay between sustainable fashion trends, social media, and second-hand shopping behaviors. Through qualita

tive 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.

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 Jr, J. F., Venkatesh, A., Wang, J., & Chan, R. (2012). Fast fashion, sustainability, and the ethical appeal of luxury brands. Fashion Theory 16(3), 273-296. https://doi.org/10.2752/175174112X13340749707123
2. Kim, A. J., & Ko, E. (2012). Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. Journal of Business Research 65(10), 1480-1486. https://doi.org/10.1016/j.jbusres.2011.10.014
3. Caro, F., Martínez-de-Albéniz, V. (2015). Fast Fashion: Business Model Overview and Research Opportunities. In: Agrawal, N., Smith, S. (eds) Retail Supply Chain Management. International Series in Operations Research & Management Science, vol 223. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7562-1_9
4. Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., ... Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. International Journal of Information Management 59, 102168. https://doi.org/10.1016/j.ijinfomgt.2020.102168
5. Johnstone, L., Lindh, C. (2022). Sustainably sustaining (online) fashion consumption: Using influencers to promote sustainable (un)planned behaviour in Europe's millennials. Journal of Retailing and Consumer Services 64, 102775. https://doi.org/10.1016/j.jretconser.2021.102775
6. de Oliveira, L. G., Miranda, F. G., de Paula Dias, M. A. (2022). Sustainable practices in slow and fast fashion stores: What does the customer perceive? Cleaner Engineering and Technology 6, 100413. https://doi.org/10.1016/j.clet.2022.100413
7. Persson, O., Hinton, J. B. (2023). Second-hand clothing markets and a just circular economy? Exploring the role of business forms and profit. Journal of Cleaner Production 390, 136139. https://doi.org/10.1016/j.jclepro.2023.136139
8. Palomo-Domínguez, I., Elías-Zambrano, R., & Álvarez-Rodríguez, V. (2023). Gen Z's Motivations towards Sustainable Fashion and Eco-Friendly Brand Attributes: The Case of Vinted. Sustainability 15(11), 8753. https://doi.org/10.3390/su15118753

THE NEXUS BETWEEN ARTIFICIAL CHATBOTS AND INDIVIDUAL CREATIVITY

Luana Coelho[1]
Orlando Lima Rua[2] [0000-0002-1593-7440]

Keywords: Artificial Intelligence, Chatbots, Individual Creativity, SEM.

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]. [4] argues that 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. These models are used in a variety of applications, such as chatbots, virtual assistants, language translation, and text generation.

Literature Review

[5] define chatbot as a sophisticated computer program designed to simulate human conversation and interact intelligently with users. 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] argue that it empowers chatbots to understand the nu-

1 Porto School of Accouting and Business (ISCAP), Polytechnic Institute of Porto (P.PORTO), Rua Jaime Lopes Amorim, s/n, 4465-004 S. Mamede de Infesta, Portugal.
2 Porto School of Accouting and Business (ISCAP), Polytechnic Institute of Porto (P.PORTO), Rua Jaime Lopes Amorim, s/n, 4465-004 S. Mamede de Infesta, Portugal; 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; Research Center of Business Sciences (NECE), University of Beira Interior (UBI), 6200-209 Covilhã, Portugal, lncs@springer.com

ances 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.

Individual creativity is a fundamental human capacity that enables people to generate novel and valuable ideas, solutions, and products [9]. According to [10], 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 [11].

Methodology

Data was collected from a students' online survey of the Porto School of Accouting 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 [12]. Chatbot usage intention and performance expectancy (CB) were examined using three items each [13]. Compatibility (C) was measured with the adoption of six items [14]. 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 [15], with subsequent modifications introduced by [16], and [17]. 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).

Results

PLS-SEM was used to test the hypotheses with SmartPLS 3.0 software [18]. 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 (β=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 (β=0.121, t=1.589); thus, H2 was supported. Moreover, C had a significant and positive effect on IC (β=0.492, t=6.119); thus, H3 also was supported. Table 2 and Figure 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.

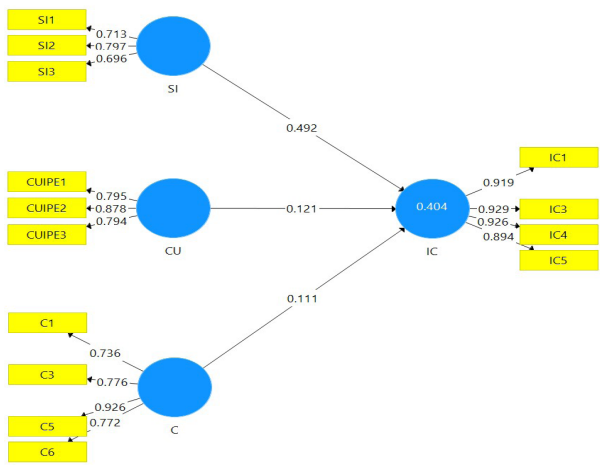


Fig. 1. Structural model assessment.

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 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.

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

References

1. Goldberg, Y. (2019). A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research 57, 345-420.
2. Dumitriu, D., & Popescu, M. A. (2020). Artificial Intelligence Solutions for Digital Marketing. Procedia Manufacturing 46, 630-636.
3. Jurafsky, D., & Martin, J.H. (2019). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Pearson Education.
4. Manning, C. D. (2020). Computational linguistics and deep learning. Annual Review of Linguistics 6, 365-387.

5. Barnes, S. J., & Vidgen, R. T. (2018). The impact of chatbots on customer service: A systematic review. Journal of Service Management 29(2), 233-256.
6. Khanna, Anirudh & Pandey, Bishwajeet & Vashishta, Kushagra & Kalia, Kartik & Bhale, Pradeepkumar & Das, Teerath. (2015). A Study of Today's A.I. through Chatbots and Rediscovery of Machine Intelligence. International Journal of u- and e-Service, Science and Technology 8, 277-284.
7. Adamopoulou, E., & Moussiades, L. (2020). An Overview of Chatbot Technology. Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II, 584, 373–383. Ramesh, V., & Singh, A. (2017). Chatbots: A comprehensive overview. Journal of Big Data 4(1), 1-22.
8. Hughes, D. J., Lee, A., Tian, A. W., Newman, A., & Legood, A. (2018). Leadership, creativity, and innovation: A critical review and practical recommendations. The Leadership Quarterly 29(5), 549–569.
9. Torrance, E. P. (2002). Why teach thinking? Creativity and giftedness. Gifted Child Today 25(3), 48-57.
10. Mumford, M. D., & Gustafson, S. B. (1988). Creativity syndrome: Integration, application, and innovation. Psychological Bulletin 103(1), 27-43.
11. Oliveira, T., Faria, M., Thomas, M.A., & Popovic, A. (2014). Extending the understanding of mobile banking adoption: when UTAUT meets TTF and ITM. International Journal of Information Management 34(5), 689-703.
12. Venkatesh, V., Morris, M.G., Davis, G.B., & Davis, F.D. (2003). User acceptance of information technology: toward a unified view. MIS Quarterly 27(3), 425-478.
13. Moore, G.C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. Information Systems Research 2(3), 192-222.
14. Tierney, P., Farmer, S.M., & Graen, G.B. (1999). An examination of leadership and employee creativity: the relevance of traits and relationships. Pers. Psychol 52, 591–620.
15. Farmer, S.M., Tierney, P., & Kung-Mcintyre, K. (2003). Employee creativity in Taiwan: an application of role identity theory. Acad. Manag. J. 46(5), 618–630.
16. Lin, C., Wittmer, J. L. S., & Luo, X. (2022). Cultivating proactive information security behavior and individual creativity: The role of human relations culture and IT use governance. Information & Management 59(6), 103650.
17. Hair, J., Hult, G.T.M., Ringle, C., Sarstedt, M. (2016). A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). London, UK: Sage.

WHITE PAGE

DIGITAL IDENTITY AND CONTROL: HOW AI REPLICAS CHALLENGE PERFORMANCE RIGHTS

AHMED QAYYUM[1], STACY A. DOORE[2] [0000-0002-8322-6798]

Extended Abstract

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 billions of dollars and highlighted how AI integration threatens traditional definitions of authorship and performance rights. This paper examines the ethical implications of AI in entertainment production, focusing on the philosophical questions raised by digital replicas. Through analysis of recent industry developments and fictional narratives, we explore how AI challenges concepts of performer autonomy through digitally replicated performers and AI generated narratives. The stakes in this discussion extend beyond individual artistic rights to the future of creative expression for a broader society.

Background

While early entertainment 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 and handling post-production tasks [1]. Major studios have reported efficiency gains of up to 40% in certain production phases by means of AI implementation [2]. However, entry-level positions for creative talent face replacement through digital automation, disrupting established career pathways. The implications of this shift go far beyond job displacement, threatening the foundations of fair compensation and creative control in an AI-driven entertainment industry [3]. The concentration of technological power in the hands of a small number of major studios raise concerns about market diversity and creative independence.

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 [4]. Digital replica technology enables the creation of synthetic performances that replicate an actor's likeness, voice, and mannerisms. Generating new performances from existing data creates a "simulacrum", a copy without an original,

1 DEPT. OF COMPUTER SCIENCE, COLBY COLLEGE, WATERVILLE ME, USA
2 DEPT. OF COMPUTER SCIENCE, COLBY COLLEGE, WATERVILLE ME, USA SADOORE@COLBY.EDU

challenging traditional notions of performance authenticity and artistic ownership [5]. A key point of contention centers on "synthetic performers"- AI-generated performances not based on specific actors - versus "digital replicas" of actual performers, reflecting complex relationships between identity, control, and commercial exploitation.

Fictional Narratives and Industry Practices

The entertainment industry's grappling with digital replicas highlights key ethical concerns through both speculative and real-world cases. The Black Mirror episode "Joan is Awful" [6] depicts a woman whose life is adapted into an AIgenerated series, without her meaningful consent. The story centers on the exploitation of personal data by a tech giant to produce inexpensive, hyper-realistic content. The film "The Congress" [7] is a prescient exploration of digital replica ethics about an actor who sells her digital likeness to a studio, allowing them to create new performances without her involvement. The film predicted many of the ethical issues now facing the industry around performer autonomy and the commodification of identity [8].

The digital recreation of the deceased Peter Cushing in "Rogue One" (2016) [9] raised questions about posthumous performance rights and the boundaries of estate consent. While Cushing's estate approved the performance recreation, the case sparked conversations about whether such approval adequately addresses the ethical implications of digitally resurrecting performers [10]. Recent films have used AI technologies to create younger versions of actors, raising questions about control of one's historical likeness. 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 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.

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 [11]. 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 [12].

Building upon a Kantian deontological perspective, digital replicas violate the categorical imperative by treating performers as means rather than ends in themselves [10]. Infinitely reproducing and manipulating an actor's likeness without ongoing consent reduces their autonomy to mere instrumental value. A traditional utilitarian might counter

by citing benefits such as preserving actor performances for future generations. These benefits must be weighed against potential harms, including the psychological impact on performers abdicating control over their own body and autonomy, the erosion of professional opportunities and legal protections, and the broader societal implications of normalizing synthetic performances [13]. Moreover, the commercial deployment of digital replicas raises questions about power dynamics and consent that extend beyond the entertainment industry. While contracts may provide legal permission for digital reproduction, the temporal nature of AI advancement means individuals cannot truly provide informed consent for future uses that may not be technologically possible at the time of agreement. This creates a 'consent paradox' where individuals must make decisions about their digital rights without truly understanding potential future implications [4].

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. New contractual frameworks should include explicit provisions for continuous consent, requiring studios to obtain renewed permission for each novel use of a digital replica. 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 [14]. 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. New legal frameworks should expand publicity rights to cover AI-generated content, including visual reproductions, synthesized voices, mannerisms, and performance styles [11]. Policies should address the use of creative works in AI training data and establish industry-wide standards for licensing creative content for AI development. This would help protect creative work while providing clear frameworks for fair compensation in an evolving technological landscape. 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 [15].

Based on the 2023 entertainment labor negotiations, actors now have the right to bargain over future use of synthetic performers and will be fairly compensated for the time spent to create their own digital replicas [14]. We argue that this same move towards 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 contributions of this paper provide a foundation for future research on ethical creative sector AI integration. However, the successful implementation of these type of responsible AI protections requires an active commitment from all entertainment industry stakeholders: studios, creative professionals, technology developers, and industry regulatory bodies.

References

1. F. Luchen and L. Zhongwei, "Chatgpt begins: A reflection on the involvement of AI in the creation of film and television scripts," *Frontiers in Art Research*, vol. 5, no. 17, 2023. [Online]. Available: https://doi.org/10.25236/FAR.2023.051701

2. J. G. Kim, "Current use and issues of generative ai in the film industry," *Journal of Information Technology Applications and Management*, vol. 31, no. 3, pp. 181–192, 2024. [Online]. Available: https://doi.org/10.21219/jitam.2024.31.3.181
3. J. Gordon-Levitt, "If artificial intelligence uses your work, it should pay you," 2023, washington Post. [Online]. Available: https://www.washingtonpost.com/opinions/2023/07/26/joseph-gordonlevitt-artificial-intelligence-residuals/
4. A. Yamazaki, "Digital replicas and democracy: issues raised by the hollywoodactors' strike," *Humanities and Social Sciences Communications*, vol. 11, p. 1661, 2024. [Online]. Available: https://doi.org/10.1057/s41599-024-04204-w 5.
5. J. Baudrillard, *Simulacra and Simulation*. University of Michigan Press, 1994.
6. C. Brooker and A. Pankiw, "Joan is Awful (black mirror, season 6, episode 1)," 2023, netflix.
7. A. Folman, "The congress," 2013, pandora Filmproduktion.
8. C. James, "The congress: The prescient film that foretold hollywood's ai crisis,"2023, bBC. [Online]. Available: https://www.bbc.com/culture/article/20230717how-2013-film-the-congress-predicted-hollywoods-current-ai-crisis
9. G. Edwards, "Rogue one: A star wars story," 2016, lucasfilm, Walt Disney Motion Pictures.
10. A. Preminger and M. B. Kugler, "The right of publicity can save actors from deepfake armageddon," *Berkeley Technology Law Journal*, vol. 39, no. 2, 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4563774
11. A. Curren, "Digital replicas: Harm caused by actors' digital twins and hope provided by the right of publicity," *Texas Law Review*, vol. 102, p. 155, 2023. [Online]. Available: https://texaslawreview.org/wpcontent/uploads/2024/01/Curren.Printer.pdf
12. Y. Wang, "Synthetic realities in the digital age: Navigating the opportunities andchallenges of ai-generated content," 2023, authorea Preprints. [Online]. Available: https://www.techrxiv.org/doi/full/10.36227/techrxiv.23968311.v1
13. S. Bender, "Generative-ai, the media industries, and the disappearance ofhuman creative labour," *Media Practice and Education*, pp. 1–18, 2024. [Online]. Available: https://doi.org/10.1080/25741136.2024.2355597
14. K. Shetler, "AI and consent: What the SAG-AFTRA and WGA agreementstell us about the future of generative AI," 2024. [Online]. Available: https://scholarship.shu.edu/student_scholarship/1483/
15. L. Kavitha, "Copyright challenges in the artificial intelligence revolution: Transforming the film industry from script to screen," *Trinity Law Review*, vol. 4, no. 1, pp. 1–8, 2023. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/104027427/1_copyright_challenges_in_the_18-_libre.pdf

SYSTEMATIC REVIEW ON AI ETHICS IN PRIVACY FOR V2X COMMUNICATIONS

Héctor Dave Orrillo Ascama[1] [0000-0003-1555-6821], Laércio Cruvinel Júnior[2] [0000-0002-7672-3243], Mário Marques da Silva[3] [0000-0002-5498-3220]

Keywords: Blockchain, Machine Learning, Federated Learning, Autonomous Vehicles, V2X communications, AI Ethics.

Extended abstract

Introduction

The advancement of autonomous vehicles (AVs) introduces significant potential for urban mobility; however, their vulnerability to cyber threats poses serious concerns. Ensuring privacy and integrity of data exchanged via vehicle-to-everything (V2X) communication is a crucial challenge (Shahriar et al., 2024; Orrillo et al., 2024). AVs collec vast amounts of sensitive data through V2X networks, sensor systems, and onboard units, making them potential targets for cyberattacks (Wang et al., 2024). Ethical concerns related to data usage, transparency, and accountability further complicate the implementation of AI-based strategies in AVs systems (Poszler et al., 2021). Addressing these challenges requires robust security measures and adherence to ethical principles (Kumar et al., 2024).

The present systematic review explores existing strategies for safeguarding AV data, focusing on the integration of Blockchain and Federated Learning technologies. These solutions aim to enhance security and ethical compliance in AVs systems. Our study presents guidelines for ensuring responsible data handling in AVs while addressing ethical considerations in AI-based strategies. The integration of AI ethics is essential to address the challenges of privacy, security, and data processing in Avs. The objective of this study is to investigate AI-driven strategies to protect data privacy and integrity in

1 Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa, 1169-023 Lisboa, Portugal; Autónoma TechLab, 1169-023 Lisboa, Portugal; Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal; horrillo@autonoma.pt

2 Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa, 1169-023 Lisboa, Portugal; Autónoma TechLab, 1169-023 Lisboa, Portugal; Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal; lcruvinel@autonoma.pt

3 Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa, 1169-023 Lisboa, Portugal; Autónoma TechLab, 1169-023 Lisboa, Portugal; Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal; mmsilva@autonoma.pt

AVs, with a particular focus on ethical considerations in V2X communications. Specifically, the study seeks to:

- Analyze existing literature on AI strategies for AVs cybersecurity.
- Identify gaps in current implementations concerning AI ethics.
- Propose a framework that integrates Blockchain and Federated Learning to enhance the privacy of data exchanged between Avs.

Methodology

This research was conducted following the PRISMA guidelines for systematic reviews. Renowned databases such as IEEE Xplore, SpringerLink, and Scopus were used. The final selection of articles for the review was determined through a structured two-step process:

1° Initial Search

A total of 8,000 articles were obtained using combinations of terms and Boolean operators. The search terms used were Data privacy AND V2X communication AND Blockchain AND autonomous vehicles AND AI Ethics.

2° Application of Filters

To refine the literature selection, we applied filters to focus on the most relevant and recent studies. The following filtering steps were undertaken:

- Limit the search results to the past 3 years (2022 to 2025); the authors obtained 1,800 documents.
- Limit the search results to peer-reviewed articles and reviews. The authors obtained 170 documents.
- Use advanced search features, such as searching for articles with the highest number of citations, subject-specific filters, and full-text search resources. The authors selected the 64 most relevant and suitable articles.

The specific filters applied to select articles were:

For privacy strategies considering AI ethics: (“data privacy” OR “data protection”) AND (“Artificial Intelligence” OR “machine learning”) AND (“autonomous vehicles” OR “AVs”) AND (“AI ethics” OR “privacy” OR “data governance”).

For cybersecurity in V2X considering AI ethics: (“cybersecurity” OR “machine learning”) AND (“V2X”) AND (“autonomous vehicles” OR “AVs”) AND (“AI ethics” OR “privacy” OR “data governance”).

For blockchain in vehicular communications with AI ethics: (“Blockchain”) AND (“Federated Learning” OR “machine learning”) AND (“V2X” OR “vehicular communication”) AND (“AI ethics” OR “privacy” OR “data governance”).

For cybersecurity in vehicular communications with AI ethics: ("Federated Learning" OR "distributed learning") AND ("GANs" OR "Deep Learning" OR "Reinforcement Learning") AND ("autonomous vehicles" OR "AVs") AND ("AI ethics" OR "privacy" OR "data governance").

Subsequently, the authors conducted the literature review process, which began with abstract screening, where the abstracts were analyzed to assess their relevance. Following this, the selected articles were read in greater depth to better understand their methods, results, and conclusions. Citation tracking was also considered, following references in more significant works to select the most specific articles on the topic at hand.

Preliminary results

The review identified several strategies aimed at enhancing data privacy and integrity in AVs. However, most approaches lack sufficient integration of ethical principles, such as transparency and accountability (Park et al., 2024). The proposed integration of Blockchain and Federated Learning offers several advantages, including:

- Data Privacy: Ensuring secure and anonymous data sharing among AVs.
- Security: Preventing unauthorized data access and manipulation.
- Decentralization: Reducing reliance on centralized servers, enhancing resilience.

Moreover, ethical issues related to data handling, including user consent, equity, and data ownership, must be duly considered. This paper identifies technical gaps in current implementations regarding the integration of ethical principles in AI. To address these gaps in the context of AVs, the following guidelines are proposed: It is essential that users are clearly informed about the data that will be collected by the On-Board Unit (OBU) of the AV, thus ensuring informed consent. Additionally, it is recommended that AVs includes mechanisms that allow users to access, modify, or delete their data, thereby promoting control and transparency.

Furthermore, it is suggested that personal data be stored and transmitted with appropriate security guarantees, using encryption techniques and effective access control policies. The importance of limiting data retention to the strictly necessary period, as previously defined, is also emphasized. Additionally, the adoption of anonymization techniques that safeguard the identity of individuals is recommended. Finally, it is argued that data should not be reused for purposes other than those originally intended within the scope of V2X communications, in accordance with the principle of purpose limitation.

Conclusion

This work highlights the advances and challenges in the use of AI to ensure data privacy and integrity in V2X communications, with a focus on key ethical principles. The integration of Blockchain (BL) and Federated Learning (FL) offers benefits such as in-

creased security and decentralization, but faces challenges, such as high computational costs and latency, which need to be addressed to make these solutions viable on a scale.

This study contributes to consolidating knowledge on autonomous vehicles, identifying gaps in research, and proposing guidelines that integrate technology and ethics. The analysis indicates that, despite technological approaches focusing on security and privacy, the lack of integration of ethical principles, such as transparency, fairness, and effective governance, is a critical flaw. Companies must adopt robust ethical governance frameworks that take these dimensions into account to protect data, promote transparency, and ensure accountability, thereby securing public trust.

Ethics in AI is fundamental, especially in contexts involving large volumes of sensitive data and interactions with autonomous vehicles, as these factors make governance processes more challenging and complex. The implementation of ethical practices is not only a technical necessity but also a social imperative, aiming to balance technological advancements with individual rights and data protection.

The contributions of this study are essential for advancing academic knowledge on AI ethics in autonomous vehicles and guiding companies in the creation of systems that respect ethical principles, ensuring security, privacy, user trust, and regulatory compliance in the digital future. Implementing frameworks that align technology, and ethics is crucial to ensuring that AI systems meet the standards required by modern society, mitigating risks, and promoting more responsible and sustainable development.

References

1. Shahriar, M. H., Ansari, M. R., Monteuuis, J. P., Chen, C., Petit, J., Hou, Y. T., & Lou, W. (2024). Vehigan:Generative Adversarial Networks for Adversarially Robust V2X Misbehavior Detection Systems. Proceedings - International Conference on Distributed Computing Systems, 1294–1305. https://doi.org/10.1109/ICDCS60910.2024.00122
2. Orrillo, H., Sabino, A., & Marques da Silva, M. (2024). Evaluation of Radio Access Protocols for V2X in 6G Scenario-Based Models. Future Internet, 16(6).
3. https://doi.org/10.3390/fi16060203
4. Wang, H., Qi, W., Jia, J., & Xing, Y. (2024). Federated Learning for V2X Communications: Opportunities and Challenges. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB.
5. https://doi.org/10.1109/BMSB62888.2024.10608239
6. Poszler, F., & Geislinger, M. (2021). AI and Autonomous Driving: Key ethical considerations. https://ieai.mcts.tum.de/
7. Kumar Engineering Manager, A., & Deere, J. (2024). Exploring Ethical Considerations in AI-driven Autonomous Vehicles: Balancing Safety and Privacy. In Issue 1 Journal of Artificial Intelligence General Science (Vol. 2). JAIGS.
8. http://creativecommons.org/licenses/by/4.0
9. Park, S., Kim, D., & Lee, S. (2024). Enhancing V2X Security Through Combined Rule-Based and DL-Based Local Misbehavior Detection in Roadside Units. IEEE Open Journal of Intelligent Transportation Systems.
10. https://doi.org/10.1109/OJITS.2024.3479716

SYMBOLIC ASPECTS OF ONLINE PRIVACY PROTECTION BEHAVIOUR: FROM A SOCIAL COMMUNICATION PERSPECTIVE

Yasunori Fukuta[1] [0000-0002-2712-5538], Kiyoshi Murata[2] [0000-0002-5632-8078],
Yohko Orito[3] [0000-0002-4293-1722], Vanessa Bracamonte[4] [0009-0007-3844-5856],
Takamasa Isohara[5] [0000-0003-2371-3698]

Keywords: Privacy Protection Behaviour, Symbolic Aspects of Privacy Protection Behaviour, Social Communication, Privacy Paradox, Social Media Users.

Extended Abstract

This study aims to explore the symbolic nature of behaviours exhibited by social media users to protect their own and others' privacy. Previous research has positioned privacy protection behaviour (PPB) as a rational action, determined by the perceived magnitude of threats and the effectiveness of countermeasures in mitigating privacy risks [1]. However, studies on the privacy paradox suggest that actual protective actions frequently exhibit systematic deviations from what would be considered rational behaviour [2]. Drawing on the well-established interactionist perspective, which posits that human behaviour functions as a form of communication, conveying messages to others in social contexts [3][4], this study investigates the symbolic aspects of PPBs through semi-structured interviews and a large-scale questionnaire survey. These symbolic aspects may help explain the systematic deviations from rational action observed in social media-based PPBs, thereby offering valuable insights into the privacy paradox. Furthermore, they may provide valuable implications for the development of legal frameworks, technologies, and services related to privacy protection.

A substantial body of research has examined PPBs and related strategies. For example, Young and Quan-Haase [5] investigated privacy protection strategies employed by university students on Facebook, such as excluding contact information, using limited profiles, untagging or deleting photos, and restricting friend requests from strangers. Their findings highlighted how these measures were employed to address threats to social privacy. Similarly, Baruh et al. [6] identified measures including managing privacy settings, using anti-spyware tools, and employing anonymity software or encryption. Their meta-analysis of 44 prior studies demonstrated a significant positive correlation between PPBs and privacy concerns.

1 Meiji University, Chiyoda, Tokyo 101-8301, Japan, yasufkt@meiji.ac.jp
2 Meiji University, Chiyoda, Tokyo 101-8301, Japan
3 Ehime University, Matsuyama, Ehime 790-8577, Japan
4 KDDI Research, Inc., Fujimino, Saitama 356-8502, Japan
5 KDDI Research, Inc., Fujimino, Saitama 356-8502, Japan

A commonly employed theoretical framework for explaining PPB is Protection Motivation Theory (PMT) [1][7]. Originally proposed by Rogers [8] and subsequently expanded, PMT conceptualises decision-making in PPB as a function of threat appraisal (e.g., severity and likelihood) and coping appraisal (e.g., efficacy of coping strategies and self-efficacy). In this view, PPB is understood as rational action taken to mitigate privacy risks, based on the perceived magnitude of the threat and the effectiveness of protective measures.

However, PPBs enacted by social media users can be seen as embedded within the context of social interaction, akin to the communicative practices that are typical of these platforms. Drawing on Herbert Blumer's theory of symbolic interactionism [3] and Erving Goffman's dramaturgical approach [4][9], such behaviours can be understood as part of interactions that occur in settings where individuals are aware of their social roles and perform expected behaviours and attitudes in the presence of others. In such situations, even when individuals perceive privacy risks, they may refrain from engaging in protective behaviours due to concerns about the impressions these actions might generate. Social media users making privacy-related decisions are often strongly motivated to construct online identities, manage their reputations, and project a desired self-image through the information they share [7][9][10]. Accordingly, it is crucial to consider PPB as embedded in the dynamics of social communication.

In this study, we defined explicit PPBs as those recognisable as protective actions by communication counterparts (e.g., editing faces or backgrounds in images before posting) and analysed how such behaviours were recognised within actual social media communication, based on findings from a semi-structured interview study and a questionnaire survey.

Semi-structured interviews were conducted via Zoom with nine university students in Japan between October and December 2024. Respondents were first asked about their use of social media, followed by questions concerning their perceptions and feelings when posting or sharing photos that have been edited to obscure or alter facial features or backgrounds, or when encountering such photos shared by others.

The interviews revealed several key insights. First, social media users appear highly aware of the communicative implications of their PPB. For instance, Participant A (male, 20s, student) explained that he used to edit photos (e.g., by adding stickers or mosaics to faces) on a private account during high school, but has since stopped. He noted, "Since people around me stopped editing photos, it feels awkward to edit them now. … Everyone else stopped, so I thought I shouldn't do it either." This suggests that the practice of editing photos for privacy protection is influenced by how such actions align with social norms and how they are perceived by others. Participant B (male, 20s, student), who primarily uses his social media account for interactions with friends, responded as follows when asked about his reaction to receiving photos with privacy edits: "I can't help but wonder why they edited it. It feels like they're overly concerned, and I don't want to be seen that way." While his awareness of privacy issues may be limited, it is significant that he interprets such edits as communicative signals—in this case, suggesting "excessive concern".

Second, the symbolic aspects of PPBs appear more pronounced in limited-access spaces, such as private SNS accounts. Participant C (female, 20s, student) noted,

"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." Participant D (male, 20s, student) also emphasised trust in closed communication spaces, stating, "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. It's about mutual trust."

Finally, the symbolic dimension of explicit PPBs was frequently associated with negative connotations, such as appearing "overly concerned with privacy" or implying "a lack of trust in others". From the perspective of impression management, individuals may avoid engaging in explicit PPBs to prevent being perceived negatively. In this sense, symbolic considerations may consciously or unconsciously discourage social media users from adopting explicit PPBs, potentially leading to inadequate privacy protection.

To examine whether these findings were observed among a broader population and to assess their generalisability, a questionnaire survey was conducted. This survey was administered between 12 and 20 April and targeted university students at Meiji University, Ehime University, and Toyama University in Japan. Out of a total of 748 responses, 508 were included in the analysis after excluding those that did not meet the criteria for social media usage or were deemed inappropriate upon screening. Respondents were asked questions regarding, among other topics, their perceptions of explicit PPBs, such as posting privacy-edited images with blurred or sticker-covered faces or backgrounds on a private Instagram account.

While the full details of the survey findings are presented in the full paper, it is worth noting that the key insights derived from the semi-structured interviews were largely corroborated by the results of the large-scale questionnaire. More than 75% of respondents acknowledged the necessity and effectiveness of such explicit PPBs as privacy-protective measures. At the same time, 43.90% expressed concern that engaging in these behaviours might make them appear overly concerned about privacy, and 40.94% felt that such practices might discourage others from posting or sharing unedited photos.

These findings suggest that, although the utility of PPBs for privacy protection is clearly recognised, their symbolic dimension—namely, the impact such behaviours may have on how others perceive or respond to them—is also evident. Notably, only 21.06% of respondents agreed that these PPBs improve self-presentation, while 43.90% agreed with the view that such actions might give the impression of being excessively privacy-conscious. This points to a potential concern that the adoption of explicit PPBs could negatively affect how one is perceived by others.

On the other hand, only around 20% of respondents agreed with the statements that such behaviours might cause discomfort to others or imply a lack of trust. Based on these findings, it may be concluded that, while these PPBs are not viewed as fundamentally damaging to one's image, they are perceived as somewhat awkward or socially difficult to perform, indicating a subtle yet meaningful barrier to implementation.

In conclusion, while prior research on privacy self-management has emphasised rational decision-making based on threat and coping evaluations, the results of this semi-structured interview study and the questionnaire survey suggest that social media users' recognition of PPB is deeply embedded in social communication. PPBs can function as a form of communicative expression. Although the primary function of such behaviours

is to mitigate privacy risks, their symbolic aspects also play a significant role, offering a multifaceted perspective on PPB. Additionally, the observed concerns that such explicit PPBs might negatively affect self-presentation suggest that these symbolic aspects may consciously or unconsciously inhibit SNS users' engagement in PPBs, potentially resulting in insufficient privacy protection. Developing a multifaceted understanding of PPB is expected to offer both theoretical insights for privacy paradox research and practical implications for fields such as SNS management, privacy policy development, and the development of privacy-enhancing technologies.

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.

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

References

1. 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. Decision Support Systems 135. (2020) 113323
2. Barth, S., De Jong, M.D.: The privacy paradox–investigating discrepancies between expressed privacy concerns and actual online behavior–a systematic literature review. Telematics and Informatics 34(7), 1038-1058, (2017)
3. Blumer, Herbert.: Symbolic interactionism: Perspective and method. Prentice-Hall, Inc., (1969)
4. Goffman, E.: The presentation of self in everyday life. Doubleday & Company Inc. (1959)
5. Young, A. L., Quan-Haase, A.: Privacy protection strategies on Facebook: the internet privacy paradox revisited. Information, Communication & Society, 16(4), 479-500. (2013)
6. Baruh, L., Secinti, E., Cemalcilar, Z.: Online privacy concerns and privacy management: A meta-analytical review. Journal of Communication 67(1), 26-53 (2017)
7. Adhikari, K., Panda, R. K.: Users' information privacy concerns and privacy protection behaviors in social networks. Journal of Global Marketing, 31(2), 96-110. (2018)
8. Rogers, R. W.: A protection motivation theory of fear appeals and attitude change1. The Journal of Psychology 91(1), 93-114. (1975)
9. Goffman, E.: Behavior in public places: Notes on the social organization of gatherings. The Free Press of Glencoe. (1963)
10. Marwick, A. E., Boyd, D.: Networked privacy: How teenagers negotiate context in social media. New media & society, 16(7), 1051-1067. (2014).

IMPLEMENTING AI INTO ENGINEERING TECHNOLOGY AND COMPUTER SCIENCE COURSES

Richard Cozzens[1]
Cathy Chang[2]

Keywords: Artificial intelligence, AI, ChatGPT

Extended Abstract

Introduction

Artificial Intelligence (AI), particularly ChatGPT, has significantly impacted the educational community. The use of AI and the ethical questions related to it has been a major topic of discussion and concern in the faculty senate at Southern Utah University (SUU), it became particularly relevant to my teaching when a student used AI to enhance the efficiency of a design in one of my engineering courses. This experience highlighted the immediate potential of AI to transform education, particularly in STEM and related educational fields. This motivated me to explore its applications in curriculum development, learning, and assessment. Recognizing both the opportunities and challenges AI presents, I decided to examine its role in STEM-related education.

AI can be integrated into nearly every aspect of education. For instance, Mollick and Mollick (2023) outline five strategies for using AI to improve teaching effectiveness, including its use as a teaching assistant. However, the literature that particularly captured my attention included Amanda Hirsch's (2024) "The Digital Red Pen: Efficiency, Ethics, and AI Assistance" and Rahul Kumar's (2023) "Faculty Members' Use of AI to Grade Student Papers: A Case of Implications." These studies underscore the potential of AI to streamline grading processes, providing faster, more consistent feedback to student's crucial benefits in education. Motivated by these insights, I sought to integrate AI into my engineering courses and collaborated with Professor Chang in her computer science courses to broaden the scope of its impact. Although there are various strategies for implementing AI to improve teaching and learning, we focused on using AI to enhance efficiency, ensure consistent grading, and provide detailed, timely feedback to students. The primary objective of this paper is to present our initial strategies for implementing AI in grading and providing timely feedback to students, as well as to document the outcomes and lessons learned from this initial implementation. While this approach may appear straightforward to experienced AI users, my experience over the past

1 Southern Utah University, UT 84721, USA, cozzens@suu.edu
2 Southern Utah University, UT 84721, USA

year has revealed that many of my colleagues remain unfamiliar or even apprehensive about the role of AI in education. It is this audience who are just beginning to explore the potential of AI in their teaching practices—that I hope to reach by documenting our experience of implanting AI at a basic level.

Methods

My foundational approach to curriculum development has been guided by established frameworks, including Butcher & Wilson-Strydom's "A Guide to Quality in Online Learning" (2012), the Quality Matters (QM) rubric, and the Roblyer (2010) rubric. These frameworks emphasize structured, student-centered learning environments and rigorous grading standards. I have applied these concepts across various teaching formats, including face-to-face, hybrid, and distance learning for both rural high schools and international students in Wuhan, China.

Given that this was the first semester applying AI to engineering and computer science courses, we decided to use observational data, applying a framework similar to my previous curriculum development strategies, which emphasize iterative improvement through Double Loop Learning (Batista, 2006). Rather than gathering extensive quantitative data at this stage, we prioritized real-world application, observing results, evaluating outcomes, and making iterative adjustments. This approach has proven effective in my past curriculum development experiences, including publishing two books, developing eight different state-level Pre-Engineering curricula, and conducting several international presentations. These experiences included addressing unique challenges, such as providing opportunities for students in small high schools in rural Utah and developing a two-plus-two program between SUU and Wuhan Polytechnic University (WPU) in Wuhan, China. This program presented new challenges, including language barriers and cultural differences, which required innovative curriculum adjustments, such as bilingual eBooks and hybrid instructional methods.

Building on these experiences, I collaborated with Professor Chang from the Computer Science department at SUU. Combining my background in STEM-related curriculum development with Professor Chang's computer science expertise, we worked together to integrate AI tools, primarily ChatGPT, into our curricula. I began by using scoring rubrics to assignments in my CCET 4610 Advanced Projects course, while Professor Chang integrated similar rubrics into her CSCY 2400 Technology and Ethics course. Each rubric was generated by ChatGPT and exported into Canvas LMS. Although our original plan included broader AI integration across multiple courses using the five strategies mentioned earlier, time constraints limited our initial efforts to this narrower scope.

Results

Our initial observations supported the literature cited by Hirsch and Kumar, confirming that AI can significantly reduce grading time and enhance feedback quality. However, we also found that AI is far from a perfect or fully autonomous solution. Although we applied AI to most of the assignments, the first few were used to practice and refine our process, again applying Double Loop Learning to improve efficiency.

Professor Chang observed that ChatGPT, when used for AI-assisted grading, sometimes allowed previously graded papers to influence the evaluation of current submissions, despite all papers being assessed against the same rubric. This raised concerns about consistency and potential bias in the grading process. In this instance, the reports were all on the same subject, potentially contributing to the issue. In contrast, my class involved individual student projects on different topics, where this bias was less apparent.

Despite these challenges, we found the overall process significantly more efficient, roughly cutting our usual grading time by one-third. This included a thorough review of the AI-generated results. The bullet points generated by AI for student feedback proved particularly valuable, as they could be quickly copied into the Canvas LMS note section, providing students with clear explanations of their grades. The consistency of this feedback, even when the rubric was applied multiple times to the same report, was a notable improvement over traditional manual grading.

Discussion on Ethics and Privacy

The use of AI in education, particularly for automated grading, involves critical ethical considerations, including data privacy, academic integrity, and legal implications. In preparation for implementing AI in our courses, we reviewed the relevant institutional guidelines at Southern Utah University (SUU). While no specific policy on AI use in the classroom currently exists, SUU has adopted "Guiding Principles for Generative Artificial Intelligence," adapted from the Responsible AI Manifesto for Marketing and Business. These guidelines do not explicitly restrict AI use in education, providing flexibility for innovation.

To align with these principles, we prioritized Responsible AI Development, Academic Freedom, and Accountability. We transparently informed students that AI would be used to generate grades and feedback, creating opportunities for them to share their perspectives on this approach. To address privacy concerns, we removed personal identifiers from each student report before processing with AI, consistent with the SUU guidelines emphasizing legal and privacy protections.

While ethics and privacy remain ongoing challenges in adopting AI in education, our experience shows that these concerns can be effectively managed through clear communication, thoughtful policy alignment, and responsible technology use.

Conclusion

Automating routine grading tasks with AI allows instructors to focus on more impactful, personalized interactions with students, supporting deeper learning and engagement. Based on our initial experiences, we found that most students are receptive to AI-assisted grading, provided that:

1. The professor is transparent about the use of AI in grading.
2. Students understand how AI will influence their grades.
3. Personal information is protected throughout the grading process.

Students appreciated the quick, detailed feedback generated by AI, which aligns well with the broader trend of AI adoption in the engineering and computer science industries. Exposing students to AI in an ethical, responsible manner better prepares them for real-world professional environments.

Using AI for grading and feedback is just one application, but it serves as a practical starting point for educators exploring the broader potential of AI in education.

References

1. Argyris, C. "Theories of action, double loop learning and organizational learning." Retrieved from http://infed.org/thinkers/argyris.htm, accessed 01.23.2011.
2. Batista, E. (2006). "Double-Loop Learning and Executive Coaching." Retrieved from http://www.edbatista.com/2006/12/doubleloop_lear.html, accessed 01.23.2011.
3. Butcher, Neil and Wilson-Strydom, Merridy. A Guide to Quality in Online Learning. s.l. : Academic Partnerships, 2012. https://www.qualitymatters.org/rubric. Quality Matters Rubric. [Online] [Cited: February 21, 2015.]
4. Garrison, D., 2007. Online Community of Inquiry Review: Social, Cognitive, and Teaching Presence Issues. Journal of Asynchronous Learning Networks, 11(n1), pp. 61-72.
5. Hirch, Amanda, The digital red pen: Efficiency, ethics, and AI-assisted grading, Northern Illinois University Center for Innovative Teaching and Learning, Oct. 29, 2024 https://citl.news.niu.edu/2024/10/29/the-digital-red-pen-efficiency-ethics-and-ai-assisted-grading/
6. Kumar, R. Faculty members' use of artificial intelligence to grade student papers: a case of implications. International Journal for Educational Integrity 19, 9 (2023). https://doi.org/10.1007/s40979-023-00130-7
7. Roblyers, M.D and Wiencke, W.R., 2010. Design and Use of a Rubric to Assess and Encourage Interactive Qualities in Distance Courses. American Journal of Distance Education, 12:2, 77-98.

SOME PROBLEMS IN THE ETHICAL IMPACT ASSESSMENT OF EMERGING TECHNOLOGIES AND SOCIO-TECHNICAL VISIONS: CASE CITYVERSE

Antero Karvonen[1], Jaana Leikas[2], Nina Wessberg[3], Anton Sigfrids[4], Sanni Pöntinen[5]

Keywords: Metaverse ethics, smart city, metaverse, participatory impact assessment

Introduction

As MacIntyre [1] recognized in discussing the emergence of philosophical ethics in ancient Greece, in periods of social and moral transformation the concepts used in framing moral questions no longer appear clear and unambiguous. Clearly, the contemporary world is in a transitional period – geopolitical, technological, environmental, and social. Consequently, the moral order is undergoing subtle (and less subtle) changes.

Technology is transforming rapidly, seemingly expanding reality's dimensions. The metaverse presents a space where established social systems and norms aren't fully formed, challenging even basic moral frameworks. While actions like virtual violence don't fundamentally challenge core ethics, they complicate their application in this digital realm. Computer culture's subversive nature sometimes fosters creativity while offering escape from reality and stress [2]. The metaverse represents a natural progression in our ability to reinvent ourselves digitally, blurring online and offline boundaries, delocalizing us, while increasing potential interactions in the infosphere [3]. This convergence may trigger a legitimacy crisis [4] where traditional understanding of human culture no longer adequately frames evolving forms of life. Philosophical and participatory foresight should guide our understanding of the evolving relationship between technology and values [6].

The acceptability and desirability of the metaverse has been subject for debate [6, 7]. For instance, the EU Parliament [8] has published a report identifying concerns about metaverse applications. These include concerns regarding health, inappropriate behaviour in the virtual world, privacy, mass surveillance, targeted surveillance and influence, and autonomy violations and manipulation. Clearly, addressing ethical concerns of metaverse use is crucial. By considering future goals, terms for achieving them, and involved values, we can better harness the metaverse's potential benefits.

1 VTT Technical Research Centre of Finland Ltd, Espoo, Finland, antero.karvonen@vtt.fi
2 VTT Technical Research Centre of Finland Ltd, Espoo, Finland
3 VTT Technical Research Centre of Finland Ltd, Espoo, Finland
4 VTT Technical Research Centre of Finland Ltd, Espoo, Finland
5 Faculty of Management and Business, Tampere University, Tampere, Finland

This paper offers a brief meditation on these matters, based on experiences gained from a semi-structured focus group method [9] used to explore perceptions of the nature and values of the CityVerse in the city of Tampere. Our focus in this paper will be less on the specific ethical contents of the findings, and more on the problems generated at the confluence of abstract socio-technical visions regarding emerging technologies and participatory ethical impact assessment.

Case Study

Forward looking cities have begun imagining the metaverse as a next step within Smart City development. The city of Tampere in Finland was among the first to formulate a Metaverse vision in 2023. Tampere's CityVerse vision - a "Cognitive City" - blends the physical and virtual domains, and facilitates immersive experiences, social interaction, learning, entertainment, and commercial activities. The vision aims to support core values of the city, including trust in technology and decision makers, community action, the well-being of individuals, sustainability, and to make the city pioneer in leadership and good governance. Visions are essentially hypothetical constructs, that articulate on an abstract level desirable directions for organizations or in this case, cities. Visions can be evaluated both from the standpoint of their contents and their use as strategic tools [10]. The present study was positioned in the latter, particularly in terms of deployment and rollout of the vision to city municipal officials and servants.

Method

A semi-structured focus group method [9] was employed to explore perceptions of the nature and values of the CityVerse. As a starting point, paradigmatic case descriptions of CityVerse for city inhabitants for 2040 were presented. The groups analyzed the CityVerse cases using values chosen collaboratively by the research team and city representatives that represented important city values. The central research question was: How can CityVerse be utilized in a manner that fosters trust, strengthens community and individual well-being, and promotes sustainability? A total of 30 city officials attended the two-hour focus groups. 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 [11].

Findings

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.

Conclusions

The ethical discussions in our findings have value for future CityVerse development cases. However, their primary significance lies in demonstrating the potential of participatory methods for implementing city-level visions. The concerns raised about insufficient justification for this technology suggest a possible legitimacy crisis [4] among city officials. While speculative, this warrants attention as citizens may develop similar legitimacy concerns about the vision.

Furthermore, the fact that more questions than answers were generated during the discussions may be taken as further evidence for this conclusion. However, the are other factors likely at play, which warrant consideration. By definition, emerging technologies are novel, and therefore there are limited concrete experiences upon which the ethical evaluations can be based. Rather, in participatory settings, they are presumably based on the subjective amalgamations of given example scenarios and whatever related experiences the subjects may bring to bear on the generated mental representations. This is issue is compounded in the case of the metaverse by the fact that it entails (for many people) a novel experience. Thus, people may generate their mental representations based on intuitions of the concrete world of experience, or perhaps existing forms

of virtual reality to guide their evaluations. This may not be altogether misguided, but throws doubt on subjects' ability to realistically share in their mental models, and thus whether they afford sufficient granularity for ethical discourse.

These challenges notwithstanding, it can be argued that nevertheless life and development must be acted forward amidst uncertainty, and therefore articulated visions can be used to guide ethical development. The problem, however, is largely the same: the visions do not, and cannot represent concrete futures at the granularity or complexity required for nuanced ethical analysis. And even if they could, the bounded rationality of human action, at least in transformations of this scale, precludes much of the nuance proper to ethical discourse during implementation.

The metaverse's potential impacts remain unclear and require further study. However, ethical examination of imaginaries shouldn't be viewed as evaluating concrete future realities. Instead, participatory impact assessment should evaluate imaginaries themselves and how they reveal participants' ethical norms. Sociotechnical visions are better understood as platforms for ethical discourse rather than concrete plans for organizational implementation. This highlights both the value and limitations of ethical impact assessment, emphasizing the importance of designing these processes around achievable goals and recognizing that the significant transformations the metaverse implies require ethical discourse. Our findings suggest technology alone cannot legitimately justify the social transformations it creates – socio-technical visions must be grounded in collective understanding of clear benefits from deployment. Participatory ethical impact assessment serves as a valuable tool for this purpose.

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. Turkle, S.: The Second Self: Computers and the Human Spirit. Granada Publishing Limited, London. (1984)
3. Floridi, L.: The Fourth Revolution: How the infosphere is reshaping human reality. Oxford University Press. (2014)
4. von Wright, G.H.: The varieties of goodness. 1st edn. Routledge (1963)
5. Vallor, S.: Technology and the virtues: A philosophical guide to a future worth wanting. Oxford University Press. (2016)
6. Chalmers, D. J.: Reality+: Virtual worlds and the problems of philosophy. Penguin UK. (2022)
7. Coeckelbergh, M.: All too real metacapitalism: towards a non-dualist political ontology of metaverse. Ethics Inf Technol 26, 30. https://doi.org/10.1007/s10676-024-09768-4 (2024).
8. European Parliament. Metaverse. JURI Committee study, Policy Department for Citizens' Rights and Constitutional Affairs, Directorate-General for Internal Policies. https://www.europarl.europa.eu/RegData/etudes/STUD/2023/751222/IPOL_STU(2023)751222_EN.pdf, last accessed 2025/01/31
9. Parker, A., Tritter, J. Focus group method and methodology: current practice and recent debate. International Journal of Research & Method in Education 29(1), 23-37 (2006)

10. Larwood, L., Falbe, C. M., Kriger, M. P., & Miesing, P.: Structure and meaning of organizational vision. Academy of Management journal, 38(3), 740-769. (1995)
11. 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. Second Edition. Bloomsbury Publishing, p. 318–329. (2017)

GAMER'S DILEMMA: INTENT MATTERS

Kai K. Kimppa[1] [0000-0002-7622-7629]
Juhani Naskali[2] [0000-0002-7559-2595]

Keywords: Gamer's Dilemma, Computer Games, Virtual Violence.

Extended Abstract

Gamer's Dilemma has been a topic of discussion since 2009 when Morgan Luck (2009) introduced it. Put shortly, the dilemma goes as follows: Why is it ok to murder in computer games, but not ok to molest children in computer games? Intuitively there seems to be something somehow more wrong about introducing child molesting into a game than there is in introducing murder into it. On the other hand, in most societies, murder is decreed to be worth a much harsher sentence than molesting a child is, due murder's unfixable nature; however hard fixing the consequences of molesting a child is, fixing having been murdered is just not possible.

Many authors have attempted a solution to the issue at hand. Both Ramirez (2020) and Hemmingsen (2024) claim, in our view correctly, that Gamer's Dilemma rests on a misunderstanding, a wrong framing of the issue. As Heinrichs (2021) points out terminological choices matter – we agree. Thus, Ali's (Ali, 2022) definitions, based on the original framing by Luck (2009)

1. Virtual murder in video games is morally permissible.
2. Virtual paedophilia in video games is not morally permissible.
3. There is no morally relevant difference between virtual murder and virtual paedophilia in video games.

Is a prime example of muddled terminological choices, wrong framing and misunderstanding of the issue following from that. First, virtual paedophilia ought to be called virtual child molestation – paedophilia is a medical condition, not something morally or legally problematic as long as it is not enacted. But more importantly, the main problem for many authors, Luck (2009) and Ali (2022) included, is that they consider virtual murder to be comparable to virtual child molestation.

Montefiore and Formosa (2023) claim that we intuitively feel that engaging in virtual murder in videogames is permissible; do we though? Nader (2020) gets close to what we argue here, but not quite there – it is not about accepting that aiming for goals

1 University of Turku, Turku School of Economics, 20014, Finland
2 University of Turku, Turku School of Economics, 20014, Finland

justifies any kind of actions, but that it justifies certain kinds of actions, and the problem with framing is still with us. None of the proposed solutions seem quite satisfactory.

On top of the previous more analytical approaches, there have been attempts to explain the dilemma through a more relativistic traditions (see e.g. Bourne & Caddick Bourne, 2019; Cecchinato, 2024; Coghlan & Cox, 2023; Kjeldgaard-Christiansen, 2020; Rouillé, 2024; Young, 2024) and from an empirical direction (see e.g. Bartel, 2012; Formosa et al., 2024; Montefiore et al., 2024). However, as the focus of this paper is in the more generalizable analytic answer, we will not go deeper into these solutions, as they do not approach the issue in a similar manner.

Our proposal for an answer stems primarily from intent and secondarily from what it tells us about the character of the players who would prefer to see child molesting in the game, rather than from consequences of the act of virtually molesting a child (if any).

Luck (2009) and others after him (e.g. Ali, 2015) seem to misinterpret the issue. They claim that murder in computer games per se is considered to be acceptable. After a more careful analysis, this does not seem to be the case. While killing for a cause can be considered morally unobjectionable, murder for murder's sake is frowned upon, almost equally to child molestation for child molestation's sake. This distinction holds true for virtual acts, as well. As an example, the game Hatred[3] received, if not the same level of criticism (e.g. see the abstract of the wiki page on the game[4]) as a game similarly focusing on child molestation, close to it. Carmageddon and the Postal series received similar condemnation and bans in several countries, due to their glorification of indiscriminate murder. If the game has no other purpose than murder or child molesting for murder's or child molesting's sake, it is increasingly difficult for the game to be accepted as healthy entertainment. Also, unlike Ali (2015) claims, there is a difference between using murder, or rather killing, as a means to an end than using child molestation as a means to an end. The latter is much harder to justify, and thus typically regresses to just doing it for its own sake. And when it does, the justification relies on intention to molest children, not on achieving something worthwhile, whereas killing typically can be justifiable, if the goals are other than just (killing or) murder for the sake of murder.

Since most games that include killing, even murder, have an internally justified reason for these actions, the relevant factor in separating acceptable and unacceptable actions in games seems to be intent rather than consequences. Killing in a game in order to save a city from destruction can feel reasonable, as the act has a believable justification within the plot. Child molesting has no such believable justification; "save the world, the only way to do this is by molesting children" would hardly be acceptable reasoning, even if it was presented as a justification within the plot. Child molestation for the greater good is not reasonable as the justification is not logically sound. Thus, the Gamer's Dilemma compares two actions that carry wildly different intents.

A clear distinction should also be drawn between the intent of the in-game character and the intent of the real-world gamer. While the former has some effect on the moral analysis, it does so only as a part of the fictional story and the presented in-

3 https://store.steampowered.com/app/341940/Hatred/
4 https://en.wikipedia.org/wiki/Hatred_(video_game)

game justification for the virtual acts. While the presented in-game justifications affect real-world intent, intent does not necessarily follow from the presentation. What we propose as the primary deciding factor in the moral analysis of virtual acts, in the context of Gamer's Dilemma, is the intent of the gamer. In many open world games, it is possible for a gamer to play them reprehensibly, taking maniacal pleasure in the slaughter of civilians merely for the sake of causing (virtual) harm. In such cases, both intuition and our analysis would place at least some moral blame on the individual gamer, not the game designers. Similarly, it is possible for a gamer to play a game like Hatred, which glorifies indiscriminate killing, for other reasons besides taking pleasure in killing, which would make the act not morally reprehensible, or at least less so.

What makes child molestation in games more reprehensible than killing in games is the gamer's intent of gaining pleasure from child abuse per se, as it is very hard to see any mitigating justifications for such acts. If the gamer's intention for abusing children in video games is to experience joy from (virtual) child abuse, this is reprehensible; just as reprehensible as killing in video games with the intention of experiencing joy from (virtual) murder. Most commonly, even in violent video games, the real-world intent of the gamer is not to derive pleasure from virtual murders, but some combination of succeeding in a challenging task, enjoying the plot and audiovisual artistry of the game, and identifying with the controlled character. This non-reprehensible intent is oftentimes supported by in-game justifications of self-preservation or working for the greater good or other rationalizable goal. Such mitigating justifications ring hollow for virtual child abuse and raise well-founded questions about the real intent of the gamer. Intent creates a real-world moral distinction between virtual murder or killing and virtual child abuse.

In conclusion, it is difficult to see murder per se being more justified than child molestation per se. However, there does seem to be a difference between using them as a means to an end. Neither is acceptable, per se, and using child molestation as means to an end suffers from lack of internal logic, raising questions about the real motivations – the intent for playing – of gamers engaging in such acts in ways that killing, or even murder, as a means to an end does not. The Gamer's Dilemma thus compares murder, when it should compare killing rather than murder, as a means to an end to child molestation, which creates a false dichotomy that resolves itself when intent is taken into account.

Disclosure of Interests: The authors have no competing interests.

References

1. Ali, R. (2015). A new solution to the gamer's dilemma. Ethics and Information Technology, 17(4), 267–274. Scopus. https://doi.org/10.1007/s10676-015-9381-x
2. Ali, R. (2022). The video gamer's dilemmas. Ethics and Information Technology, 24(2). Scopus. https://doi.org/10.1007/s10676-022-09638-x
3. Bartel, C. (2012). Resolving the gamer's dilemma. Ethics and Information Technology, 14(1), 11–16. Scopus. https://doi.org/10.1007/s10676-011-9280-8
4. Bourne, C., & Caddick Bourne, E. (2019). Players, Characters, and the Gamer's Dilemma. Journal of Aesthetics and Art Criticism, 77(2), 133–143. Scopus. https://doi.org/10.1111/jaac.12634
5. Cecchinato, M. (2024). Double-Standard Moralism: Why We Can Be More Permissive Within Our Imagination. British Journal of Aesthetics, 64(1), 67–87. Scopus. https://doi.org/10.1093/aesthj/ayad008

6. Coghlan, T., & Cox, D. (2023). Between death and suffering: Resolving the gamer's dilemma. Ethics and Information Technology, 25(3). Scopus. https://doi.org/10.1007/s10676-023-09711-z
7. Formosa, P., Montefiore, T., Ghasemi, O., & McEwan, M. (2024). An empirical investigation of the Gamer's Dilemma: A mixed methods study of whether the dilemma exists. Behaviour and Information Technology, 43(3), 571–589. Scopus. https://doi.org/10.1080/0144929X.2023.2178837
8. Heinrichs, J.-H. (2021). Virtual action. Ethics and Information Technology, 23(3), 317–330. Scopus. https://doi.org/10.1007/s10676-020-09574-8
9. Hemmingsen, M. (2024). Framing the Gamer's Dilemma. Ethics and Information Technology, 26(3). Scopus. https://doi.org/10.1007/s10676-024-09798-y
10. Kjeldgaard-Christiansen, J. (2020). Splintering the gamer's dilemma: Moral intuitions, motivational assumptions, and action prototypes. Ethics and Information Technology, 22(1), 93–102. Scopus. https://doi.org/10.1007/s10676-019-09518-x
11. Luck, M. (2009). The gamer's dilemma: An analysis of the arguments for the moral distinction between virtual murder and virtual paedophilia. Ethics and Information Technology, 11(1), 31–36. Scopus. https://doi.org/10.1007/s10676-008-9168-4
12. Montefiore, T., & Formosa, P. (2023). Crossing the Fictional Line: Moral Graveness, the Gamer's Dilemma, and the Paradox of Fictionally Going Too Far. Philosophy and Technology, 36(3). Scopus. https://doi.org/10.1007/s13347-023-00660-5
13. Montefiore, T., Formosa, P., & Polito, V. (2024). Extending the Gamer's Dilemma: Empirically investigating the paradox of fictionally going too far across media. Philosophical Psychology. Scopus. https://doi.org/10.1080/09515089.2024.2354432
14. Nader, K. (2020). Virtual competitions and the gamer's dilemma. Ethics and Information Technology, 22(3), 239–245. Scopus. https://doi.org/10.1007/s10676-020-09532-4
15. Ramirez, E. J. (2020). How to (dis)solve the Gamer's Dilemma. Ethical Theory and Moral Practice, 23(1), 141–161. Scopus. https://doi.org/10.1007/s10677-019-10049-z
16. Rouillé, L. (2024). Ludic resistance: A new solution to the gamer's paradox. Ethics and Information Technology, 26(2). Scopus. https://doi.org/10.1007/s10676-024-09772-8
17. Young, G. (2024). The gamer's dilemma: An expressivist response. Ethics and Information Technology, 26(2). Scopus. https://doi.org/10.1007/s10676-024-09766-6

EPISTEMIC INJUSTICE IN AI-DRIVEN HEALTHCARE

HALEH ASGARINIA[1] [0000-0003-2344-4060]

Keywords: AI-Driven Healthcare, Epistemic Injustice, Explainable AI, Transparency.

Abstract

Artificial Intelligence (AI) is increasingly integrated into healthcare, with predictive models being used for early diagnosis, treatment recommendations, and patient monitoring. One such application is using AI models trained on voice and patch sensor data to predict Chronic Obstructive Pulmonary Disease (COPD) and provide lifestyle improvement suggestions[2]. While AI systems promise enhanced efficiency and accuracy, their integration into medical decision-making introduces risks of epistemic injustice. Drawing on the frameworks provided by Fricker (2007), and Carel and Kidd (2014), this paper explores the epistemic injustices that may arise for both patients and practitioners when AI models operate without adequate transparency and justification mechanisms.

Epistemic injustice, as conceptualized by Fricker (2007), includes testimonial injustice—where certain knowers are systematically discredited—and hermeneutical injustice—where individuals lack the conceptual tools to make sense of their experiences. In the context of AI-driven COPD prediction, testimonial injustice may manifest when patients' lived experiences and concerns are overridden by AI-driven predictions, diminishing their role as credible knowers of their health. This is particularly concerning for marginalized groups whose voices have historically been excluded or devalued in medical discourse. If a patient's voice data does not align with AI-driven assessments, they may be unable to challenge the AI's prediction due to a lack of epistemic authority. Additionally, hermeneutical injustice arises when patients, particularly those from disadvantaged or non-dominant groups, do not have access to or understanding of how AI-driven decisions are made, leaving them without the resources to question or interpret their diagnosis effectively.

For medical practitioners, epistemic injustice materializes in two critical ways: (1) their professional judgment may be undermined if the AI model's predictions are prioritized over their expertise, leading to a shift in epistemic authority from human professionals to opaque algorithms, and (2) their decision-making may be constrained by a lack of justification for AI-driven outputs, making it difficult to critically engage with and contest the AI's predictions. This is particularly problematic when practitioners interact

1 MAASTRICHT UNIVERSITY, NETHERLANDS, HALEH.ASGARINIA@MAASTRICHTUNIVERSITY.NL

2 THIS EXAMPLE REFERS TO THE DACIL PROJECT. FOR MORE INFORMATION, SEE: HTTPS://WWW.NWO.NL/EN/PROJECTS/KICH1GZ0321023

with AI systems that operate as "black boxes," providing results without explainability. Without insight into the AI's reasoning, practitioners may either be forced to accept its outputs blindly or struggle to defend their clinical assessments when they diverge from AI-driven outputs. This can lead to cognitive overload, reduced professional autonomy, and epistemic exclusion, reinforcing systemic biases in healthcare decision-making.

The harms associated with these epistemic injustices extend beyond individual interactions to structural issues within the healthcare system. Patients may experience increased mistrust in medical practitioners and AI-assisted healthcare, leading to disengagement and poorer health outcomes. Practitioners, on the other hand, may face professional disempowerment, exacerbating stress and dissatisfaction in medical practice. Additionally, the lack of epistemic transparency in AI decision-making perpetuates power asymmetries, wherein AI developers and institutions hold epistemic authority while patients and healthcare providers are positioned as passive recipients of AI recommendations.

To mitigate these epistemic injustices, several measures should be implemented. First, AI models should be designed with explainability and transparency, providing justifications for their outputs in language accessible to both practitioners and patients. This ensures that both groups can critically assess AI-driven recommendations rather than purely deferring to them. Second, AI systems should incorporate feedback loops where practitioners can contest and refine AI decisions based on clinical expertise, allowing for dynamic model adjustments. This approach aligns with epistemic justice by ensuring that AI is not treated as an infallible authority but rather as a tool that assists human decisions. Third, patient-centered AI governance should prioritize epistemic inclusivity by involving diverse patient populations in model training, validation, and oversight. This would reduce the likelihood of marginalized groups experiencing testimonial or hermeneutical injustice due to their unique physiological or sociocultural factors being inadequately represented in training data.

Furthermore, institutional policies should emphasize the role of epistemic humility in AI-assisted healthcare, recognizing the limitations of both AI and human decision-making. By fostering collaborative epistemic environments where AI recommendations are subject to professional scrutiny and patient input, healthcare systems can uphold fairness and accountability in medical knowledge production. Finally, educational programs for practitioners should integrate AI literacy training, equipping them with the skills to interpret, question, and contextualize AI predictions, thereby reinforcing their epistemic agency in clinical decision-making.

In conclusion, while AI models offer significant advancements in COPD prediction and healthcare delivery, their deployment must be critically examined through the lens of epistemic justice. Without deliberate interventions, AI can perpetuate testimonial and hermeneutical injustices, silencing patients and eroding the professional authority of practitioners. By prioritizing transparency, practitioner feedback, and inclusive model design, healthcare institutions can mitigate these risks and foster a more equitable, just, and epistemically responsible AI-driven healthcare system.

Acknowledgement: This research was conducted as part of the DACIL project at Maastricht University, funded by the Dutch Research Council (NWO).

References

1. Carel, H., & Kidd, I. J. (2014). Epistemic injustice in healthcare: A philosophical analysis. Medicine, Health Care, and Philosophy, 17(4), 529–540. https://doi.org/10.1007/s11019-014-9560-2
2. Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.

DESIGNING FREEDOM: ESG INVESTMENT, INDUSTRIE 4.0, BLOCKCHAIN, AND THE DYNAMICS OF NETWORK VALUE

KAZUYUKI SHIMIZU[1] [0000-0002-4847-0197]

Keywords: ESG investment, Industrie 4.0, Blockchain, Decentralised Governance, Markov chain.

Abstract

In this paper, we explore new directions for value creation in the digital economy and the construction of a sustainable society through the integration of ESG investment, Industry 4.0, and blockchain technology, focusing on the following three points.

First, based on Max Weber's theory, investment is conceptualized as a process of self-realization, representing an ethical transformation that goes beyond the traditional capitalism model focused on monetary returns, by incorporating elements of Environment, Social, and Governance (ESG).

Second, we examine advances in network infrastructure within Industry 4.0 technologies, particularly in the United States, with an emphasis on the Internet of Things (IoT) and data protection, along with the democratization of data structures through blockchain. Special attention is given to the evolution of financial networks symbolized by cryptocurrencies and to initiatives that integrate individuals' multidimensional value indicators into market mechanisms. These efforts propose building a new economic foundation that reflects not only financial indicators but also emotions and diverse values.

Finally, just as the taste of a delicious pudding varies depending on multiple factors—such as whether it is chilled in a refrigerator or made with more eggs—the cointegration relationship between Bitcoin price, time, and other variables suggests that Bitcoin prices follow a certain long-term growth trajectory over time. This indicates that traditional understandings of network structures can be reinterpreted from flexible and diverse perspectives, and that such analysis becomes possible through quantum computation by computers.

Defining ESG Investment

Investment is the allocation of capital (funds, time, effort, knowledge, etc.) to specific goods or services with the expectation of increasing value. Max Weber, in The Protestant Ethic and the Spirit of Capitalism, conceptualized investment as an ex-pres-

1 MEIJI UNIVERSITY, TOKYO 1018301, JAPAN, SHIMIZUK@MEIJI.AC.JP

sion of free will. He argued that Protestant savings, through the process of capital accumulation, were reinvested in promising industries, thereby promoting economic growth. This perspective opposed the doctrine of predestination and positioned in-vestment as a mechanism for self-realization. This process reflects how abstract ideas are actualized and how industrial capital and individual investors are enabled to par-ticipate partially in enterprises through equity investment.

This conception of investment reveals similar ethical transformations in contexts such as Islamic Sharia law, the Christian Methodist movement, and, in the contem-porary context, the development of ESG, SRI (Socially Responsible Investment), and EA (Effective Altruism). Furthermore, individuals engaged in investment through new technologies are increasingly capable of realizing highly imaginative ventures by utilizing Augmented Reality (AR) and Virtual Reality (VR).

Investment behavior is influenced by risk tolerance, and by emphasizing low-risk, high-return options, the self-replication of financial capital becomes the foundation of capitalism. However, the profits and losses derived from such investments place disproportionate recognition on the monetary value associated with personal trans-formation and virtual participation, while neglecting the broader physical dimensions of sustainability that support the natural environment (Shimizu, 2018).

By incorporating the elements of Environment, Social, and Governance (ESG) into the framework of investment, the interdisciplinary scope of investment is expanded. The environmental (E) dimension has shifted from CO□ reduction toward energy transition and carbon neutrality; the social (S) dimension has broadened from work-ing conditions to encompass a wide range of stakeholders, including employees, cus-tomers, and local communities; and the governance (G) dimension has evolved from corporate governance to the dynamics of networks that guide society as a whole. As a result, investment has advanced into a framework that evaluates not only “finan-cial returns” but also “social returns.” Ensuring sustainability requires institutional and managerial design of systems that enable smooth energy circulation, as well as the construction of stress-free feedback loops (Mitali , 2017).

Thus, in defining ESG investment, it is necessary to examine the historical develop-ment of financial networks within social structures. The conceptual expansion of financial networks, in contrast to ESG, is rooted in Socially Responsible Investment (SRI), which emphasizes individual morality and social responsibility.

Integration of Industrie 4.0 Technologies into the Network Infrastructure

Blockchain technology has revolutionized financial transactions, corporate govern-ance, and decentralized decision-making. The transition from Proof-of-Work (PoW) to Proof-of-Stake (PoS) consensus mechanisms considers improving energy efficien-cy while maintaining system autonomy. Decentralized Finance (DeFi) promotes Bitcoin as “digital gold” and Ethereum as a smart contract platform, enhancing scalability, business feasibility, and security, thereby contributing to the expansion of financial systems and investment opportunities (Shimizu, 2024).

In connecting the themes of ESG investment and Industry 4.0, particular emphasis is placed on the shared element of “data.” Data can be formalized as “data =

numbers + information (meaning)." Essentially, data represent qualitative or quantitative vari-ables belonging to a set of items, embodying the attributes or characteristics of an object. It is important to note that data inherently contain both qualitative aspects (e.g., "beautiful") and quantitative aspects (e.g., "10 meters long"). In ESG invest-ment, data that express ethical dimensions are managed at Layer 1 within data spac-es in Industry 4.0 to ensure privacy and public accountability, while at Layer 2, cor-porate actors engage in socially market-oriented competition over data utilization.

Cryptocurrencies derived from Bitcoin have assigned monetary value to networks through blockchain technology. Their value stems from network effects, exemplified by the dynamics of social media, and is compressed into financial networks through tamper-proof, fungible tokens. Under competitive conditions, such data achieve legit-imacy and highly secure trust via consensus mechanisms, offering a new source of hope distinct from the stagnation of closed systems dominated by Big Tech (GAFA, etc.). This new hope transforms formerly closed data structures and contributes to the formation of more democratic and creative data architectures, thereby "designing freedom" (Zhuming & Chris, 2025).

At the same time, discomfort arises from the reality that value standards embedded within social institutions and governance structures rely excessively on monetary value. Exploring how to reconcile institutional frameworks with diverse individual values and sensibilities—unaccepted by conventional systems—presents a compel-ling challenge. In particular, blockchain technology, Decentralized Finance (DeFi), and AI-operated Decentralized Exchanges (DEXs) are increasingly recognized for their potential to re-position individual values and capabilities as tradable assets with-in market mechanisms. Through these mechanisms, individuals' multidimensional value indicators can be dynamically adjusted and evaluated in the market, enabling the emergence of a new economic foundation and value system that transcends the traditional money-centered paradigm.

Contradictions between Time and Price, and Network Structures

Bitcoin's price fluctuations are often described as being too volatile to serve as an investment asset, or as manifestations of "black swan" events and long-tail phenom-ena. However, when Bitcoin prices are analyzed on a logarithmic scale, a long-term correlation between price and time becomes evident (see the regression line in Fig. 1 below) (Burger & Vijn, 2024). This can be attributed to the fact that, as Barabási suggested, network effects theoretically enable unlimited expansion (Albert, 2018).

If the value dynamics of a network succeed, its distribution does not follow a Gaussi-an model but instead resembles the patterns predicted by a generalized central limit theorem—patterns in which a single large-scale entropy shift fundamentally trans-forms the entire system, as observed in earthquakes or black swan events (Dániel , Márton , István , & Gábor , 2014). However, in today's digital ecosystems, the frag-mentation of knowledge is accelerating, resulting in frequent disruptive changes. Google's PageRank algorithm incorporates a Markov chain, achieving efficiency by discarding unnecessary historical data. In probability theory, Markov demonstrated that random events ultimately converge to a normal distribution regardless of tem-poral sequence [Gilder, 2018].

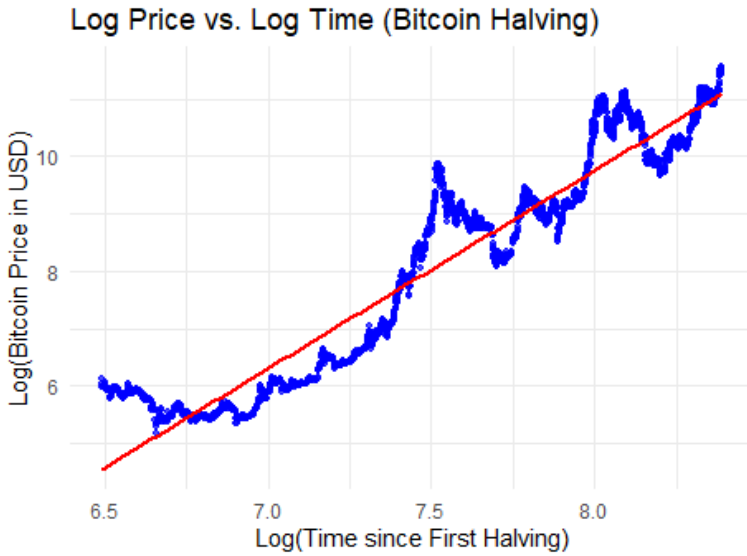


Fig. 1. BTC-USD Close Price (Log Scale).

Gilder further argues that time itself is increasingly being tokenized as a form of cur-rency, and that information constitutes the primary source of value (Gilder , 2023). When time is understood as a relative concept, Gilder suggests that the tokenization of time as currency, and the creation of value from the information embedded in time, can become a central issue.

If such diverse values can be tokenized, then the ethical returns of ESG investment and the values defined by other networks become interchangeable. In practical terms, however, Feynman pointed out that quantum mechanics is probabilistic (wave functions), and because the state space expands exponentially, an exact simulation on classical computers is infeasible (Richard, 1981).

References

Albert, B. -L. (2018). The Formula: The Universal Laws of Success. Little, Brown and Company. Retrieved from https://www.academia.edu/75141375/Albert_L%C3%A1szlo_Barab%C3%A1si_The_Formula_The_Universal_Laws_of_Success_The_Science_Behind_Why_People_Succeed_or_Fail

Alex , S. (2023, 11 4). Bitcoin Vs Ethereum Consensus: PoW Vs PoS Mechanisms Explained. Retrieved from https://www.doubloin.com/learn/bitcoin-vs-ethereum-consensus

Burger, H. C., & Vijn, P. (2024, Jan 31). Medium. Retrieved from Bitcoin's time-based power-law and cointegration revisited: https://medium.com/quantodian-publications/bitcoins-time-based-power-law-and-cointegration-revisited-15f212b95ea5

Dániel , K., Márton , P., István , C., & Gábor , V. (2014). Do the Rich Get Richer? An Empirical Analysis of the Bitcoin Transaction Network. doi:https://doi.org/10.1371/journal.pone.0086197

Gilder , G. (2023). Life After Capitalism: The Meaning of Wealth, the Future of the Economy, and the Time Theory of Money. Regnery Gateway.

Gilder, G. (2018). Life After Google: The Fall of Big Data and the Rise of the Blockchain Economy. Simon and Schuster.

JV , A. (2019). Self-Organising in Blockchain Infrastructures: Generativity through Shifting Objectives and Forking. IT-Universitetet i Kobenhavn. Retrieved from https://pure.itu.dk/ws/portalfiles/portal/84029150/JAIS_3rd_Revision.pdf

Mitali , B. (2017). Elements of Innovators' Fame: Social Structure, Identity and Creativity. Columbia University. Retrieved 5 1, 2018, from https://www8.gsb.columbia.edu/programs/sites/programs/files/abstracts/Banerjee_columbia_0054D_13831.pdf

Richard, F. P. (1981). Simulating Physics with Computers. Department of Physics. Pasadena, California: California Institute of Technology. Retrieved from https://s2.smu.edu/~mitch/class/5395/papers/feynman-quantum-1981.pdf

Shimizu, K. (2018). The Dilemma between "comply or explain" and SRI, ESG methodology; transitional terminology. Berlin, Germany: The 6th International Conference on Social Responsibility, Ethics and Sustainable Business. Retrieved from https://www.researchgate.net/publication/322594259_The_Dilemma_between_comply_or_explain_and_SRI_ESG_methodology_transitional_terminology

Shimizu, K. (2024). The countervailing power of AI DAOs influences value transformation; Bitcoin (POW) vs. Ethereum (POS). Ethicomp 2024, 9 - 19. Retrieved from https://data.mendeley.com/public-files/datasets/5mxd93pjt5/files/802ec3c4-a82f-439f-a614-d7ef1aa8882e/file_downloaded

Simon, R. (2023). Technology Ethics: The Ethical Digital Technology Trilogy. London: Routledge.

Zhuming, B., & Chris, Z. W. (2025). Internet of Things (IoT) for Sustaninable Mechatronic Systems and Cyber-Physical System (CPS). 2025 10th International Conference on Information and Network Technologies. Retrieved from https://www.icint.org/prog.pdf

SECURITY AND ETHICS IN THE USE OF COMPUTING TECHNOLOGIES AND THE INTERNET

Laercio Cruvinel Júnior[1] [0000-0002-7672-3243], António Cabeças[2] [0000-0001-8496-535X], Adriana Lopes Fernandes[3] [0000-0001-7025-9470]

Keywords: Cybersecurity, Ethical AI, Data Privacy, Algorithmic Bias, Digital Ethics, Regulatory Frameworks

Extended Abstract

Introduction

The rapid advancement of computing technologies and the pervasive use of the Internet have transformed societies, businesses, and educational institutions. However, issues such as data privacy, algorithmic bias, cybersecurity threats, digital burnout, and ethical AI applications have become central to discussions in academia, industry, and policy-making (Stahl et al., 2023; Zostant & Chataut, 2023).

This paper offers a critical synthesis of contemporary ethical and security challenges in computing, arguing that digital well-being and user empowerment must be central to policy reform and technological design. It also identifies best practices for the responsible use of computing technologies, aiming to balance innovation with ethical accountability, data protection, and digital well-being.

Ethical Considerations in Computing Technologies

Privacy, Data Protection, and Digital Well-being

Privacy is one of the most pressing ethical issues in computing technologies (Zostant & Chataut, 2023). The collection, storage, and processing of personal data raise concerns about user autonomy and the right to control one's information. The challenge is to implement transparent policies that ensure user awareness and control over their digital footprint (van Baalen, 2018).

In addition to privacy concerns, digital burnout has emerged as a major ethical issue (Pinto, 2023). Although commonly discussed as separate issues, digital burnout and algorithmic surveillance often co-occur in educational and workplace settings, revealing a broader ethical concern with techno-governance over personal autonomy.

1 Universidade Autónoma de Lisboa Luís de Camões, 1169-023 Lisboa, Portugal; Autónoma TechLab, 1169-023 Lisboa, Portugal, lcruvinel@autonoma.pt

2 Universidade Autónoma de Lisboa Luís de Camões, 1169-023 Lisboa, Portugal; acabecas@autonoma.pt

3 Universidade Autónoma de Lisboa Luís de Camões, 1169-023 Lisboa, Portugal; Escola Superior Náutica Infante Dom Henrique, 2770-058 Paço de Arcos, adrianafernandes@enautica.pt

Algorithmic Bias and Fairness

AI and machine learning algorithms increasingly influence decision-making in critical sectors such as healthcare, finance, and education (Stahl et al., 2023). However, these systems are vulnerable to biases embedded in their training data, which can lead to discriminatory outcomes that disproportionately affect marginalized populations. Addressing these issues demands ethical scrutiny.

From a deontological perspective, systems must be designed to treat all individuals with equal respect, regardless of statistical utility. In contrast, a consequentialist approach might justify certain algorithmic trade-offs if the overall benefit outweighs individual harms, such as increased efficiency or predictive accuracy.

The reasoning above can be ethically fraught when it normalizes systemic inequities under the guise of optimization. Ethical AI development should be grounded in frameworks that emphasize transparency, accountability, and inclusivity. Without such safeguards, the deployment of AI risks reinforcing existing social hierarchies rather than promoting equitable outcomes.

Autonomy, Digital Ethics, and Work-Life Balance

AI-driven recommendations and automated decisions can subtly, or overtly, diminish user influence and discernment—particularly when algorithms operate as "black boxes" with limited transparency or explainability (Knight, 2024). From a Kantian (deontological) perspective, such opacity is ethically problematic: individuals must be treated as ends in themselves, not as means to statistical or institutional efficiency. When users are subjected to decisions they cannot understand, challenge, or opt out of, their moral agency is compromised. In contrast, a utilitarian viewpoint might argue that automation improves outcomes overall, streamlining processes and reducing human error. However, this must be weighed against the psychological and social costs of reduced control and depersonalized interaction. To reconcile these tensions, ethical guidelines should prioritize explainability, informed consent, and design for user empowerment.

Moreover, prolonged engagement with digital systems contributes to digital burnout, a condition linked to stress, fatigue, and a deteriorating work-life balance (Pinto, 2023). Ethical technology design must therefore support human-centered practices, such as workload modulation, screen-time awareness, and proactive disengagement policies that respect the limits of cognitive and emotional labor.

Cybersecurity, Ethical Hacking, and Digital Exhaustion

Ethical hacking plays a critical role in safeguarding digital systems (Wang, 2023). However, the ethical implications of cybersecurity practices must be carefully examined, particularly as AI-driven surveillance tools become more pervasive.

From a consequentialist view, continuous monitoring may be justified if it prevents harm. Yet, a deontological perspective emphasizes the intrinsic value of individual privacy and autonomy, and AI-powered systems that monitor user behavior to detect threats may

unintentionally erode trust and contribute to digital exhaustion (Jawad, 2024).
Ethical cybersecurity should prioritize transparency, proportionality, and user well-being. Systems must be designed to minimize invasiveness while maintaining effectiveness, and users should be made aware of how their data is monitored and protected.

Security Threats in Computing Technologies and the Internet

Cybercrime exploits vulnerabilities in digital infrastructures, often targeting individuals, institutions, and critical systems (Shillair et al., 2015). In response, organizations have embraced advanced cybersecurity strategies such as end-to-end encryption, multi-factor authentication, and continuous network monitoring. These measures also introduce complex ethical trade-offs related to privacy, autonomy, and psychological well-being.

Machine learning algorithms can detect anomalies, predict vulnerabilities, and automate threat response mechanisms at a scale unmanageable by human analysts (Wang, 2023). Yet, the use of AI introduces its own risks—most notably, adversarial AI attacks and limited transparency, raising concerns about accountability and user agency.

From a consequentialist perspective, the use of AI in cybersecurity can be justified by the greater good it achieves: preventing data breaches, protecting infrastructure, and maintaining public safety. But this rationale becomes ethically precarious when it fails to address the deontological obligation to respect user rights. For instance, pervasive surveillance technologies can infringe on privacy, chill free expression, and foster distrust among users (van Baalen, 2018). The normalization of such practices undermines the foundational democratic value of privacy as a prerequisite for free thought and expression. Legal frameworks alone are insufficient unless paired with institutional ethics that prioritize human dignity alongside data protection (Zostant & Chataut, 2023).

Ultimately, effective cybersecurity must be both technically robust and ethically grounded, also ensuring that the tools designed to protect do not become mechanisms of control.

Proposed Solutions and Best Practices

Ethical AI Development

AI developers should adopt ethical AI principles, including:

- Transparency: Clear documentation of AI decision-making processes.
- Accountability: Mechanisms to identify and rectify biases in AI models.
- Inclusivity: Diverse and representative datasets for training AI systems.
- Human-Centered Design: AI applications that prioritize human well-being and reduce digital exhaustion.

Strengthening Cybersecurity Measures

Organizations and individuals should implement robust cybersecurity practices, such as:

- Encryption and Secure Authentication: Protecting sensitive data with strong encryption methods.
- Regular Security Audits: Conducting periodic vulnerability assessments.
- User Awareness Programs: Educating users about phishing attacks, online threats, and the impact of digital exhaustion (Shillair et al., 2015).

Regulatory and Policy Frameworks

Essential measures by governments and regulatory bodies should include:

- Data Protection Laws: Ensuring compliance with global standards.
- AI Ethics Guidelines: Establishing clear ethical standards for AI deployment.
- Consumer Rights Protection: Enforcing transparency in data collection and user consent policies.

Digital Literacy and User Responsibility

Digital literacy programs should focus on:

- Privacy Awareness: Teaching users how to safeguard their personal data.
- Critical Thinking: Encouraging skepticism towards misinformation and online manipulation.
-

While the proposed solutions aim to foster ethical and secure digital practices, their implementation is not without challenges, to be discussed in the full version of this article.

Conclusion

By considering transparency, fairness, and accountability in AI development, strengthening cybersecurity measures, enforcing regulatory frameworks, and promoting digital well-being, we can create a digital environment that upholds ethical standards and protects user rights. The future of computing technologies depends on responsible innovation that aligns with societal values, ensuring a secure, ethical, and mentally sustainable digital landscape for all. By prioritizing ethical reflexivity in design and regulation, institutions can not only enhance security but also restore public trust in digital systems—a crucial factor for democratic resilience in the information age.

References

1. Jawad, L. A. (2024). Security and Privacy in Digital Healthcare Systems: Challenges and Mitigation Strategies. Abhigyan, 42(1), 23–31. https://doi.org/10.1177/09702385241233073
2. Knight, T. D. (2024). Genealogical Ethics in the United States and the Popularization of Genealogical Research in the Digital Age. Genealogy (2313-5778), 8(3), 78. https://doi.org/10.3390/genealogy8030078
3. Pinto, M. I. F. (2023). Devem as instituições de ensino superior estar preocupadas com o digital burnout? Uma perspetiva sobre a satisfação no trabalho. Master thesis.

4. Shillair, R., Cotten, S. R., Tsai, H.-Y. S., Alhabash, S., LaRose, R., & Rifon, N. J. (2015). Online safety begins with you and me: Convincing Internet users to protect themselves. COMPUTERS IN HUMAN BEHAVIOR, 48, 199–207. https://doi.org/10.1016/j.chb.2015.01.046
5. Stahl B. C. Timmermans J. Mittelstadt B. D. (2016). The Ethics of Computing: A Survey of the Computing-Oriented Literature. ACM Computing Surveys, 48(4), 1–38. 10.1145/2871196
6. van Baalen, S. (2018). 'Google wants to know your location': The ethical challenges of fieldwork in the digital age. Research Ethics, 14(4), 1–17. https://doi.org/10.1177/1747016117750312
7. Wang, R. (2023). AI Use in Enhancing Cybersecurity for Safeguarding Digital Information. 2023 International Conference on Applied Physics and Computing (ICAPC), Applied Physics and Computing (ICAPC), 2023 International Conference on, ICAPC, 412–419. https://doi.org/10.1109/ICAPC61546.2023.00082
8. Zostant, M., & Chataut, R. (2023). Privacy in computer ethics: Navigating the digital age. Computer Science and Information Technologies; Vol 4, No 2: July 2023; 183-190 ; 2722-3221 ; 2722-323X ; 10.11591/Csit.V4i2. https://doi.org/10.11591/csit.v4i2.p183-190

HOW WILL THE MEMORY OF COVID-19 BE PASSED DOWN? THREE KEY CHALLENGES

Akiko Orita[1] [0000-0001-6004-2649]

Keywords: Pandemic Memory Preservation, Digital Legacy, Digital Remembrance

Extended Abstract

Introduction and Research Questions

The global outbreak of COVID-19 in 2020 profoundly altered daily life, particularly during the early lockdown period before the development of vaccines. Given that much of our personal records now exist in digital form, critical ethical questions arise: How will individual experiences during the COVID-19 pandemic be remembered in the long term? Who controls these collective memories, and whose narratives will shape future historical understanding?

This study addresses three main challenges in preserving the memory of COVID-19, each with significant ethical implications:

1. The distortion and fading of time perception during and after the pandemic
2. The limitations of digital archives in preserving personal experiences
3. The varying attitudes toward retaining or deleting digital records

By examining these challenges through an information ethics lens, this research addresses significant gaps in understanding how digital technologies shape collective memory of unprecedented events. While existing research has explored psychological impacts of the pandemic and digital afterlife management separately, little attention has been paid to their intersection and ethical implications for historical record-keeping.

While existing studies have explored the psychological impacts of the pandemic and digital legacy management separately, few have comprehensively addressed how these issues intersect, particularly from an information ethics perspective. This gap underlines the urgent need for ethical frameworks that consider both the reliability of memory and the preservation of diverse digital narratives.

Related Studies

Changes in Time Perception During the Pandemic

Research suggests that the lockdowns enforced worldwide in early 2020 significantly impacted people's perception of time. Biffi et al. (2021) examined how parents experienced emotional instability during lockdown[1], while Martinelli et al. (2020)

1 Kanto Gakuin University, Yokohama, JAPAN, akiko.orita3@gmail.com

found that people perceived time as moving more slowly due to boredom and reduced well-being[2]. A longitudinal study by Droit-Volet et al. (2023) revealed that two years after the initial lockdown, participants remembered time as having passed more quickly than it actually did, suggesting that certain moments had been omitted in retrospective memory[3]. These findings indicate that pandemic memory undergoes distortion, raising information ethics concerns about the reliability of digital testimonies preserved for historical purposes.

Challenges of Digital Archiving and Information Ethics

Another challenge in preserving COVID-19 memories is the reliance on digital data, which introduces significant ethical considerations regarding ownership, access, and longevity of these records. Jaime (2020) examined the security and accuracy of social media accounts belonging to deceased scholars, highlighting the need for ethical management of digital legacies[4]. Kasket (2020) argued that Big Tech companies ultimately control how and for how long these records are preserved, raising information ethics concerns that the digital traces of the pandemic may not be retained indefinitely[5].

Orita (2023) found significant cultural differences in attitudes toward social media preservation after death. In Japan, many individuals preferred to have their social media accounts deleted posthumously, whereas in the U.S., users were more inclined to retain them in memorial mode[6]. This suggests that the historical record of COVID-19 may be unevenly preserved, influenced by regional and cultural differences in digital legacy management, raising ethical issues of representation in historical documentation.

Research Methods and Key Findings

The study employed a mixed-methods approach to explore the ethical dimensions of digital memory preservation during the COVID-19 pandemic.

Research Methods

The qualitative research was conducted between November and December 2024, using semi-structured interviews with 10 Japanese residents who had elementary school-aged children in April 2020. The interviews focused on participants' recollections of the pandemic period, their social media usage, and their attitudes toward preserving digital records.

The quantitative research was conducted through an online survey of 200 Japanese residents who had lived in Japan during the 2020 state of emergency. The survey examined participants' recollections of the lockdown, social media use, and perspectives on preserving digital records, addressing gaps in our understanding of public attitudes toward digital memory ethics.

Key Findings

Distortion of Memory and Time Perception

In the qualitative interviews, none of the participants accurately recalled the actual length of the state of emergency. The quantitative survey confirmed this pattern,

with only 5% of respondents correctly remembering the duration, while 13.5% overestimated it by four months. This memory distortion raises important information ethics questions about the reliability of digital testimonies for historical preservation.

Limited Documentation of Personal Experiences

Most interviewees consumed rather than created social media content during the pandemic, using it primarily as a coping mechanism. Only one interviewee actively posted about their pandemic experiences, while the survey showed only 25% of respondents engaged in broader online discussions. This passive consumption pattern creates an information ethics challenge of representation—if most people do not document their experiences, whose narratives shape the digital historical record?

Diverse Attitudes Toward Digital Preservation

Despite not actively posting, most interviewees recognized the value of preserving social media records of the pandemic for historical purposes. Around 50% of survey respondents expressed positive views on retaining pandemic-related digital records for 50–100 years. This finding highlights the tension between individual digital rights and collective historical documentation—a key concern in information ethics.

Conclusion

The findings confirm three significant ethical challenges in preserving the memory of COVID-19. First, memory of the pandemic is not accurately retained, raising information ethics concerns about the reliability of digital testimonies. Second, most individuals were consumers rather than creators of digital content during the pandemic, creating an ethical challenge regarding whose narratives become part of the historical record. Third, the diversity of pandemic experiences and varying attitudes toward digital preservation present an ethical imperative to preserve varied narratives rather than allowing a simplified historical account to emerge.

These challenges suggest that preserving the memory of COVID-19 requires both intentional archival efforts and consideration of information ethics principles. Future research should explore frameworks for digital preservation that acknowledge memory distortion, address representation gaps, and balance individual digital rights with collective historical documentation.

This study directly addresses the core themes of information ethics by examining how digital technologies mediate collective memory, raising critical questions about ownership, accessibility, and the ethical stewardship of historical narratives in the digital age.

Acknowledgments: This research is supported by JSPS Grant-in-Aid for Scientific Research (C) #22K12724.

References

1. Biffi, E., Gambacorti-Passerini, M. B., Bianchi, D.: Parents under Lockdown: the Impacts of the COVID-19 Pandemic on Families. Rivista Italiana Di Educazione Familiare 18(1), 97–111 (2021). https://doi.org/10.36253/rief-10332

2. Martinelli, N., Gil, S., Belletier, C., Chevalère, J., Dezecache, G., Huguet, P., Droit-Volet, S.: Time and Emotion during Lockdown and the Covid-19 Epidemic: Determinants of Our Experience of Time. Frontiers in Psychology 11, 616169 (2020).
3. Droit-Volet, S., Martinelli, N., Dezecache, G., Belletier, C., Gil, S., Chevalère, J., Huguet, P.: Experience and memory of time and emotions two years after the start of the COVID-19 pandemic. PLOS ONE 18 (2023).
4. Jaime, A. T.: On the fate of social networking sites of deceased academics in the Covid-19 era and beyond. Current Research in Behavioral Sciences 1 (2020).
5. Kasket, E.: Social Media and Digital Afterlife. In: Savin-Baden, M., Mason-Robbie, V. (eds.) Digital Afterlife, pp. 21–34. CRC Press, Boca Raton (2020).
6. 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).

THE MORAL ARGUMENT FOR OPTING OUT OF INSTAGRAM, FACEBOOK, AND X

Erich Riesen[1] [0000-0003-0744-1047]

Keywords: Social media, protest, imperceptible harm

Extended Abstract

In this paper, I argue that many of us have a strong and pressing moral reason to opt out of social media platforms run by Elon Musk or Mark Zuckerberg. The basic grounds of this obligation are (1) the billionaires' increasingly erratic and immoral behaviors coupled with (2) their growing political power. I argue by analogy with imperceptible harm cases that when the moral cost is negligible and the potential good secured extremely significant, the fact that one's action is unlikely to make a difference is not morally exculpatory. In what follows, I outline the paper by discussing the immorality of imperceptible harm and marshaling evidence for (1) and (2). Note that my conclusion is conditional: If you believe that Musk and Zuckerberg's socio-political agendas are causing significant harm, then you have a reasonably strong pro tanto duty to opt out of their platforms.

Imperceptible harm cases describe situations in which individually harmless (or negligibly harmful) choices sum together to create an extremely harmful outcome (typically through collective action or inaction) [1]. Examples include voting, climate-change, factory farming, and exploitative labor. In each case, one makes a decision that has a vanishingly small likelihood of making a difference to the outcome in question. My vote won't change the election; my car won't impact global temperature; my vegan burger won't lessen animal suffering; my purchasing an alternative device won't result in less child labor. So too, my choosing a different platform won't (on its own) make the harmful effects of Musk and Zuckerberg's sociopolitical capital less likely to occur.

However, when the stakes are high, insofar as one disagrees with some end, morality bars them from supporting its means. Therefore, liberals, and more broadly, those who are anti-Trump, anti-DOGE (Department of Government Efficiency), or anti-corporatocracy have a strong, pro-tanto moral reason to disengage from platforms directly responsible for the expanding political power of Elon Musk and Mark Zuckerberg. The harm one inflicts by supporting Musk and Zuckerberg's companies may be imperceptible, but like other imperceptible harms, the fact that one's choice is, on its own, unlikely to make a difference is not morally exculpatory [6]. Ethicists largely agree on this point [1]. If there are alternative options that require sacrificing nothing of moral signifi

1 Texas A&M University, College Station TX 77802

cance while increasing the likelihood of securing a weighty moral good, then one has a pro-tanto obligation to choose them. Take factory farming as an example: Insofar as I have easy access to healthy, non-factory-farmed food, I shouldn't support the factory farming industry [6-8]. Collectively, refusing to purchase factory-farmed meat secures the good of less needless suffering. If enough of us avoid causing the imperceptible harm, our collective action makes a real difference. Our choice also sends a signal of moral disapproval to society, acting as form of nonviolent resistance to an objectionable situation.

But why think that social media use (even our collective social media use) is causally related to the obtaining of some weighty and objectionable situation? 22% of Americans use X (owned and operated by Musk). 70% of Americans use Instagram or Facebook (owned and operated by Zuckerberg)2. Each day, we spend more than 2 hours on social media [3]. Algorithms reward this interaction through dopamine hits, maximizing screen time, subtly influencing belief systems, and lining pockets. Musk's massive pockets were arguably necessary for Trump's win—the disastrous consequences of which are only beginning to materialize. And Zuckerberg has been shifting to the right for months, hoping to position himself favorably with the incoming administration [5]. By supporting Musk and Zuckerberg's companies, one is financially supporting (and tacitly endorsing) their political agendas and the agendas of people they fund (e.g., Trump). If one believes their political agendas are causing significant harm, then one has a pro tanto moral reason to abstain from interacting with the platforms that financially enable and socially cement their influence.

But why think that their political agendas are causing harm? America's techno-oligarchs seem to have stumbled down rabbit holes of their own making, with Musk increasingly buying into unfounded conspiracy theories, backing far-right nationalist parties, and distrusting the scientific expertise he once hailed as humanity's savior3. Zuckerberg, meanwhile, is reinventing himself as an anti-woke crusader for free speech (Musk already did that) [2]. Both are dismantling fact checking and content moderation, while their companies' algorithms remain conveniently hidden behind copyright and trade secret laws [6]. The concern that China is infecting American youth through TikTok is a bit rich when American social media companies already do such a good job of it, as evidenced by their platforms' role in radicalizing young men [4], spreading patently false and dangerous ideas [10], and compromising the mental health of young people [11].

Because the costs are small and the potential benefits extremely important, many of us have a pro tanto moral duty to opt-out of Facebook, Instagram, and X. Companies can be run, friends made, and art and ideas promoted on other sites. However, we can't salvage the American experiment once it fails. We can't (easily) fix the mental health of our children. And we can't (quickly) rehabilitate America's image or reclaim its moral standing. In continuing to post and comment, we are complicit in any moral harms that result from Musk and Zuckerberg's sociopolitical capital. We are complicit in their

2 These statistics come from Exploding Topics (https://explodingtopics.com/blog/x-user-stats) and Backlink (https://backlinko.com/instagram-users), respectively.

3 You can track Elon Musk's politics and power at Tech Policy Press (https://www.techpolicy.press/tracking-elon-musks-political-activities/).

disregard for truth and decency. And we are complicit in the pain they inflict on future generations.

In the paper, I respond to several objections to my argument. For example, one might claim that we can meaningfully separate a person's behavior from the products that they create. Just because Sean "P. Diddy" Combs commits sexual assault does not imply, the objection goes, that one does something morally objectionable when listening to his songs on Spotify or when purchasing a re-release of No Way Out. Assume that this is correct. Even still, the objection loses its force when applied to the case of opting out of Musk and Zuckerberg's platforms. Buying albums does not materially support P. Diddy's immoral behavior. We can meaningfully separate the two only because there is not a direct causal relationship between records sold and sexually exploitative behavior. Sean Combs' money may have enabled him to assault more high-profile victims and to get away with it for longer than he would have if he was less rich and famous. However, the odds are that Sean Combs was sexually assaulting people before the fame and would have done so regardless of whether it materialized. On the other hand, Musk and Zuckerberg would not be able to push their political agendas on the White House or the American people if they were not billionaires controlling and profiting from Facebook, Instagram, and X (the primary information ecosystems and vehicles of public debate in our society).

One might also argue that the pro tanto duty to opt out is overridden by important moral benefits provided to liberals who remain on the most popular platforms. Facebook, Instagram, and X are important tools for protest and resistance, for example. This objection loses its force, however, when one takes note of the fact that Musk and Zuckerberg's companies legally control all content on their respective sites. The First Amendment to the U.S. Constitution protects individual speech from government—not corporate—interference. Social media companies can regulate legal content in whatever way they see fit. The protest and resistance alluded to in the objection, therefore, is more likely to have its desired impact on social networking sites wherein liberals don't need to worry about their voices being suppressed, or their accounts being flagged or deleted.

Acknowledgments: I would like to acknowledge Martin Peterson for helpful comments on earlier drafts of this paper as well as the Bovay Foundation for funding my current position and paying for my trip to EthiComp 2025.

Disclosure of Interests: The author has no competing interests to disclose.

References

1. Almeida, M. J.: Imperceptible harms and benefits, vol. 8. Springer Science & Business Media (2000)
2. Goldberg, A.: Mark Zuckerberg says companies need more 'masculine energy'. What does that even mean? USA Today. https://www.usatoday.com/story/life/health-wellness/2025/01/17/mark-zuckerberg-meta-workforce-masculine-energy/77755286007/
3. Hendrix, J.: Tracking Elon Musk's politics and power. Tech Policy Press (2025)
4. Hoye, P.: Down the alt-right rabbit hole: The role of social media platforms in the radicalization of domestic right-wing terrorists. Southern California Interdisciplinary Law Journal 33(1), (2023)

5. Ingram, D.: Mark Zuckerberg's shift to the right was months in the making. NBC News. https://www.nbcnews.com/politics/mark-zuckerbergs-shift-right-months-making-politics-desk-rc-na186861
6. Norcross, A.: Puppies, pigs, and people: Eating meat and marginal cases. Philosophical Perspectives 18, 229-245 (2004)
7. Regan, T.: Empty cages: Facing the challenge of animal rights. Rowman and Littlefield (2004)
8. Singer, P.: Animal liberation: The definitive classic of the animal movement (updated). Harper Collins Publishers (2009)
9. Sun, H.: The right to know social media algorithms. Harvard Law & Policy Review 18(1), (2023)
10. Zadrozney, B.: Social media hosted a lot of fake news this year. Here is what went the most viral. NBC News. https://www.nbcnews.com/news/us-news/social-media-hosted-lot-fake-health-news-year-here-s-n1107466
11. Zubair, U., Khan, M.K. & Albashari, M.: Link between excessive social media use and psychiatric disorders. In Annals of Medicine and Surgery vol. 85(4), 875-878 (2023) https://doi.org/10.1097/MS9.0000000000000112

ETHICS OF PERSONAL DATA ECONOMY: UNPACKING OF DILEMMAS AND QUANDARIES IN MODERN DIGITAL ECONOMY

Minna M. Rantanen[1] [0000-0001-8832-5616], Anushree Luukela-Tandon[2] [0000-0002-7319-418X]

Keywords: Data economy, data ecosystems, ethics, scoping review

Extended Abstract

Data have gradually become a central element of many different sectors, making an increasing number of companies part of the data economy - a modern form of economy rooted in the use of data [1]. As the concept of data economy is still relatively new, there is no consensus about its definition, and new definitions constantly emerge [2]. Various terms, such as data ecosystem [2] and data economy ecosystem [3] describe the same thing. In particular, there are also more critical terms used to describe data economies, such as "data capitalism" [1] and "surveillance capitalism" [4]. Essentially, the data economy is an umbrella construct covering diverse phenomena in which data is collected, analysed, and utilised to create (business) value.

Data economies, supported by rapid technological advances, are evolving quickly. But, due to the complexity and multidisciplinary nature of the field, the academic research on data economies is fragmented, particularly in terms of its ethical development and applications. From a broader ethical and societal perspective, data economies offer much to study, especially if we consider personal data and how that data can be used to influence us and whole societies. Ethical data collection and use issues can be perceived as specific ones of computational operations or from a broader societal perspective. For instance, field of data ethics focuses on the ethical challenges posed by data science and its procedures [5]. Although many ethical issues concerning data and personal data are being debated in applied and practitioner circles, there is little comprehensive information regarding how ethics intersects holistically with the data economy. Thus, we argue that there is a need to examine what data economies mean in theory and practice from an ethical perspective to develop a comprehensive knowledge structure and understand the boundaries of this concept. Such knowledge is paramount for creating an ethical framework to support future growth and application of data economies.

In the full paper, we aim to comprehensively study the ethical issues of the data economy by conducting a scoping review. In this extended abstract, we present a brief overview of ethical themes handled in the literature and provide a reasonable basis for

1 University of Jyväskylä, Finland, minna.m.rantanen@jyu.fi
2 University of Eastern Finland, Finland; European forest institute, Finland

the complete study. As a starting point, a literature study on the ethical governance of data ecosystems [6] noted that ethical issues unrelated to core research ethics can be roughly divided into five categories:

- privacy
- accountability
- ownership
- accessibility
- motivations.

In addition, consent, security, trust, and transparency were frequent themes in the literature study. However, Rantanen at al. [6] note that handling ethical topics was superficial and rarely included ethical justifications [6]Although this research was limited to the governance of data ecosystems, it shows a rising interest in the importance of ethics in the data economy. Some themes occur repeatedly, thus providing a sufficient starting point. Next, we will elaborate on these themes and examine their connections to set a basis for the scoping review.

Privacy is one of the most fundamental issues concerning the collection and use of personal data. Our notions of information privacy have been changing over time [7]. However, in general, it has referred to individual's desire to control or have some influence over the data about themselves [8]. Information privacy is a popular topic that can be defined in various ways and analysed at multiple levels, ranging from individual privacy to social privacy. Privacy can be seen as a right, commodity, state, or control, and it overlaps with concepts such as confidentiality, secrecy, anonymity, and security. [9.] For instance, constant surveillance through smart devices threatens privacy, as it removes the possibility of being free from observation in places previously seen as private, such as homes [10]. Notably, people find privacy important, but still rarely try to protect their data [11].

One way to enhance privacy is through transparency, which allows individuals to control the data collection. Transparency has an evolving meaning as a concept, but it is often used as a synonym for open decision-making. It can be seen as a way to ensure organisations' accountability and counter corruption [12]Transparency is vital in controlling privacy, as informed consent is required to disclose data. Consent should be freely given and informed, and it is made possible through transparency. Therefore, transparency about the reasons for data collection and how the data is used is essential from a privacy perspective, as well as ensuring the collecting organisation's accountability.

Ethical issues concerning accountability are often seen as a question of responsibility in case of harmful events and their prevention. Moral responsibility and accountability and their relations have been widely discussed, but there is little consensus on their definitions or relations [see, e.g. , 13]. In addition, the terms are often used interchangeably. One possible differentiation is that accountability implies instrumentality and external control, while responsibility is more about morality and inner controls [14]. In other words, responsibility is more like an assigned duty, whereas responsibility is chosen.

In a data economy setting, questions about accountability and responsibility can be contemplated at the individual and organisational levels. However, it must be not-

ed that although organisations are often seen as legally responsible entities and, thus, accountable for their actions, the incumbent people (or employees) make the decisions. After all, moral agency cannot be extended to artefacts, as its idea is that an individual is primarily responsible for their intended and voluntary behaviour and therefore, artefacts cannot be considered moral agents [15].

In addition to ensuring accountability, transparency is often seen as a means to create trust. Trust essentially results from a process in which an individual evaluates trustworthiness. In the case of data economies, trust is institutional rather than personal. In institutional trust, the trustee places trust in the rules, roles, and norms of an institution instead of another human being [16]. Trust and trustworthiness play a key role in adopting and using technology [17]. In the case of personal data economies, trust towards the institutions collecting and using data is even more relevant, as it affects the individuals' decisions about enclosing personal data [18]. Also, trust between the institutions with a data economy is essential, as the whole system depends on cooperation.

Traditionally, ownership refers to the relationship between physical property and its owner. This concept becomes problematic with personal data because it is not a physical object and can be held by multiple people simultaneously. Personal data can also be used multiple times without being consumed. In the data economy, personal data is an asset that can create value for its possessor through use or sale and the individual should have rights over it, even without ownership. This dilemma can be addressed by distinguishing between legal ownership and the ethically justified control that individuals should have over their data [19].

In data economy, accessibility is a vital issue as without access data cannot be used and value created. Accessibility is about what information an individual or organisations has a right or privilege to obtain, but also about the conditions and safety of the data [20]Thus, accessibility is not only about who can use the data but also about ensuring that it can be used and that it stays secure from malicious purposes.

Rules, roles, and norms are connected to accountability and responsibility but also clarify what actions are allowed and what are not. Official rules are often enforced by penalties and require governance and supervision. In data economies, there is no hierarchical control within a system, but it does not mean there are no control mechanisms [3]. For instance, laws and policies regulate the actions of different parties. Notably, laws and policies are not inherently ethical but should be grounded in ethical considerations. Continuous technological development creates policy vacuums that could be addressed through ethical frameworks. [21]. This calls for ongoing criticality towards the policies and conceptualisations in a complex data economy ecosystem.

Motivations, as an ethical issue, require consideration of why people do what they do. For instance, Conger et al. [22] used motivation as an ethical category to raise awareness of the beneficiaries of unethical acts and those who might suffer. They divide this ethical category into 1) identifying classes of potential victims 1) rationales behind actions, and 3) direct and indirect beneficiaries. Although intentions and avoidance of harm are rather general ethical issues, they should also be considered when considering more specific ethical issues.

As seen from preceding discussion, numerous ethical issues in data economies that can be categorised in these five themes. While focusing on specific issues is nec

essary, we should also aim to understand the bigger picture of ethical problems and their inter-relationships in the personal data economy. Thus, in the full paper, we will systematically map the existing literature using scoping review guidelines by Arksey and O'Malley [23] using these themes as the basis. We find scoping review a suitable methodology as it allows us to map the currently scattered research and identify the need for further research in the field. Thus, this review would aid current and future research in data economy ethics.

Acknowledgements: This study was partly funded by Liikesivistysrahasto (grant number: 22027).

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

References

1. Sadowski, J.: When data is capital: Datafication, accumulation, and extraction. Big Data & Society. 6, 2053951718820549 (2019). https://doi.org/10.1177/2053951718820549.
2. Oliveira, M.I.S., Lima, G. de F.B., Lóscio, B.F.: Investigations into Data Ecosystems: a systematic mapping study. KAIS. 1–42 (2019).
3. Koskinen, J., Knaapi-Junnila, S., Rantanen, M.M.: What if we had fair – people-centred – data economy ecosystems? In: Proceedings of IEEE Smart World conference 2019 (2019).
4. Zuboff, S.: Big other: Surveillance Capitalism and the Prospects of an Information Civilization. Journal of Information Technology. 30, 75–89 (2015). https://doi.org/10.1057/jit.2015.5.
5. Floridi, L., Taddeo, M.: What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 374, 1–5 (2016). https://doi.org/10.1098/rsta.2016.0360.
6. Rantanen, M.M., Hyrynsalmi, S., Hyrynsalmi, S.M.: Towards Ethical Data Ecosystems: A Literature Study. In: IEEE ICE/ITMC At: Nice, France, 2019. pp. 1–9 (2019).
7. Igo, S.E.: Me and My Data. Historical Studies in the Natural Sciences. 48, 616–626 (2018).
8. Bélanger, F., Crossler, R.E.: Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems. MIS Quarterly. 35, 1017–1041 (2011). https://doi.org/10.2307/41409971.
9. Smith, H.J., Dinev, T., Xu, H.: Information Privacy Research: An Interdisciplinary Review. MIS Q. 35, 989–1016 (2011).
10. Zuboff, S.: The Age of Surveillance Capitalism. Profile Books, London, England (2019).
11. Gerber, N., Gerber, P., Volkamer, M.: Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. Comput Secur. 77, 226–261 (2018). https://doi.org/10.1016/j.cose.2018.04.002.
12. Ball, C.: What is transparency? Public Integrity. 11, 293–308 (2009).
13. Shoemaker, D.: Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. Ethics. 121, 602–632 (2011). https://doi.org/10.1086/659003.
14. Lindkvist, L., Llewellyn, S.: Accountability, responsibility and organization. Scandinavian Journal of Management. 19, 251–273 (2003). https://doi.org/10.1016/S0956-5221(02)00027-1.
15. Johnson, D.G.: Computer systems: Moral entities but not moral agents. Ethics and Information Technology. 8, 195–204 (2006). https://doi.org/10.1007/s10676-006-9111-5.
16. Smith, M.L.: Building institutional trust through e-government trustworthiness cues. Inf. Technol. People. 23, 222–246 (2010).
17. Bahmanziari, T., Pearson, J.M., Crosby, L.: Is trust important in technology adoption? A policy capturing approach. Journal of Computer Information Systems. 43, 46–54 (2003).

18. Trein, P., Varone, F.: Citizens' agreement to share personal data for public policies: trust and issue importance. Journal of European Public Policy. 31, 2483–2508 (2024). https://doi.org/10.1080/13501763.2023.2205434.
19. Koskinen, J.: Datenherrschaft–an ethically justified solution to the problem of ownership of patient information, (2016).
20. Mason, R.O.: Four Ethical Issues of the Information Age. MIS Quarterly. 10, 5–12 (1986).
21. Moor, J.H.: What is computer ethics? Metaphilosophy. 16, 266–275 (1985).
22. Conger, S., Loch, K.D., Helft, B.L.: Ethics and information technology use: a factor analysis of attitudes to computer use. Information Systems Journal. 5, 161–183 (1995). https://doi.org/10.1111/j.1365-2575.1995.tb00106.x.
23. Arksey, H., O'Malley, L.: Scoping studies: towards a methodological framework. International Journal of Social Research Methodology. 8, 19–32 (2005). https://doi.org/10.1080/1364557032000119616.

HARMS OF SMART HOMES - A SYSTEMATIC MAP

SALLY BAGHERI[1] [0009-0009-7387-3367]

Keywords: Smart Homes, Harms, Domestic Surveillance, Ethics

Abstract

In this paper, harms facilitated by smart homes are considered. A systematic mapping study is conducted, including 48 published and peer-reviewed articles. The study aims to provide an overview of the research area and support future research. Through thematic analysis, eight preliminary categories of smart home-facilitated harm are identified and discussed.

Introduction

Smart homes are domestic spaces where people live complex and idiosyncratic lives amid internet-connected devices, offering numerous benefits [1]. However, exploring harms and negative impacts of smart homes is crucial for making robust ethical evaluations. Reports of unethical and, at times, dangerous uses of these systems are increasing, with two examples being an increased level of intrusive surveillance [2] and their facilitation of domestic abuse [3]. 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 [3]). Alternatively, scoping reviews have been conducted, including risks, vulnerabilities, and users' agency in the discussion [4]. Such results tend to be too narrow (focusing on a specific harm) or too general (analyzing the types of publications and research disciplines). Therefore, this study aims to provide an overview of the current knowledge base on smart home-facilitated harm and support future research efforts. The following research question (RQ) is posed: what kinds of harms facilitated by smart homes have been identified by previous research? 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'[5]. In this sense, we are interested in identifying the forms of harm made easier or possible by smart homes.

1 DEPARTMENT OF COMPUTER SCIENCE AND MEDIA TECHNOLOGY, MALMÖ UNIVERSITY, SWEDEN, SALLY.BAGHERI@MAU.SE

EMERGE 2025 HUMAN, SOCIAL, ETHICAL IMPACTS OF ICT LISBON

Method

A systematic mapping study aims to provide a broad overview of a research area [6]. 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 [6]. The present study follows the process of conducting a systematic mapping review as suggested [6]. 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 [6], as they mentioned a particular form of harm (inclusion criterion).

Preliminary Results & Discussion

Figure 1 depicts a preliminary map of eight harms facilitated by smart homes identified in the study. The following is an initial review and discussion of some of the results underpinning the categories.

Fig. 1. A preliminary map of smart home-facilitated harm.

Aligned with previous reviews [3], domestic violence stood out among the results. The home has long been a place of violence [7], and the smart home exacerbates gender-based violence in alarming ways [8]. Smart homes have been seen to incite conflict due to increased surveillance, revealing conflict-provoking information [9]. 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 [2]; it is necessary to differentiate benign forms of surveillance from harmful ones.

Another category aligned with previous research is security breaches caused by external actors with malicious intent [10]. For example, when actors unfamiliar to the occupants manipulate devices to open a window, unlock a door [11], or cause harm remotely through malware attacks [10]. 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 [12]. Yet, users are commonly identified as the weakest link in security-related harm [10]. Although users have not

been hacked, the anxiety related to, for example, interpreting unusual device behavior as signs of being hacked causes users stress [12]. The stress caused by the uncertainty of whether one has been hacked deserves further attention.

Due to skewed power dynamics between children and their caregivers, intrusive surveillance of children is identified as a form of harm. Such practices can infringe upon children's rights to autonomy while disguised as responsible parenting [13]. The balance between care and control represents a personal decision for caregivers and a societal trajectory regarding acceptable parenting tactics [13]. Ensuring that smart home devices are child-proofed and safe also falls within this category [14].

One result discussed the harms of increased radiation due to smart home infrastructure [15]. 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 [16]. Smart homes designed for this user type risk replacing human care with data collecting and notification systems [17], 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 [16].

Big tech companies' hold on smart home markets [18] was identified as a harm. This was mainly discussed through the lens of surveillance capitalism [19] regarding how large conglomerates normalize surveillance and leverage their monopoly-like positions to stifle fair market competition [18]. As a result of their unruly access to personal data, harms related to societal and democratic harm are also discussed [18].

Two categories were directly related to domestic surveillance, contributing to the platformization of the home [19]. Phenomenologically, this leads to chilling effects; as users are tethered to the digitalized environment, they become modulated and disciplined by the smart home infrastructure [20]. As a result, users may adapt or suppress their behavior, which impedes their experience of being at home. Another category was function creep: when devices are purchased, for example, to care for one's pet (pet tech) and subsequently used for unintended purposes such as to facilitate unethical forms of surveillance [21]. Function creep is poorly understood in the context of smart homes.

Limitations

The review focused on identifying smart home-facilitated harm, and some limitations are worth noting. As mentioned, these results are preliminary, and in-depth analyses are necessary to comprehensively determine how smart homes facilitate harm. The study aimed to provide an overview of the area, contributing to supporting future research efforts. This aligns with the aim of systematic mapping studies [6]; the method prioritized breadth rather than depth in the present review. Thus, a broad and diverse range of articles have been included. In addition, although Figure 1 represents the harms as independent categories, they should be understood as layered with interconnected elements. For example, domestic abuse could interrelate to function creep, and societal and democratic harms could be sub-categories to big tech's market dominance. These examples are not explicit in the map (Figure 1), and more research is necessary

to represent such relationships accurately. A deeper examination of the literature might reveal additional insights to support ethically sound smart home living.

References

1. Lupton, D., The Internet of Things: Social dimensions. Sociology Compass, 2020. 14(4).
2. Dereymaeker, J., et al., Smarter homes, smarter surveillance? Exploring intimate surveillance practices in modern day households. New Media & Society 2024.
3. Afrouz, R., The Nature, Patterns and Consequences of Technology-Facilitated Domestic Abuse: A Scoping Review. Trauma, Violence & Abuse, 2023. 24(2): p. 913-927.
4. Olabode, S., et al., Complex online harms and the smart home: A scoping review. Future Generation Computer Systems, 2023. 149: p. 664-678.
5. Wood, M.A., et al., Inviting, affording and translating harm: Understanding the role of technological mediation in technology-facilitated violence. The British Journal of Criminology, 2023. 63(6): p. 1384-1404.
6. Petersen, K., et al. Systematic mapping studies in software engineering. in 12th International conference on evaluation and assessment in software engineering. 2008. BCS Learning & Development.
7. Slupska, J., Safe at Home : Towards a Feminist Critique of Cybersecurity. St Antony's International Review, 2019. 15(1): p. 83-100.
8. Douglas, H., et al., Technology-facilitated Domestic and Family Violence: Women's Experiences. British Journal of Criminology, 2019. 59(3): p. 551-570.
9. Del Rio, D.D.F., Smart but unfriendly: Connected home products as enablers of conflict. Technology in Society, 2022. 68.
10. Alshamsi, O., et al., Towards Securing Smart Homes: A Systematic Literature Review of Malware Detection Techniques and Recommended Prevention Approach. Information (2078-2489), 2024. 15(10): p. 631.
11. Adeeb Mansoor, A., N. Mohammed, and M. Khurram, Ontology-Based Classification and Detection of the Smart Home Automation Rules Conflicts. 2024, IEEE. p. 85072-85088.
12. Rostami, A., et al., Being Hacked: Understanding Victims' Experiences of IoT Hacking. in 18th Symposium on Usable Privacy and Security, 2022: p. 613-631.
13. Sovacool, B., et al., Can Prosuming Become Perilous? Exploring Systems of Control and Domestic Abuse in the Smart Homes of the Future. Frontiers in Energy Research, 2021. 9.
14. Sun, K., et al., Child Safety in the Smart Home: Parents' Perceptions, Needs, and Mitigation Strategies. in ACM on Human-Computer Interaction. 2021, ACM. p. 1-41.
15. Reshna, R. and A. Kheira Anissa Tabet, An Appraisal Among Wired, Hybrid and Wireless Smart Homes to Mitigate Electromagnetic Radiation, in Frontiers in Built Environment. 2022, Frontiers Media S.A.
16. Berridge, C. and A. Grigorovich, Algorithmic harms and digital ageism in the use of surveillance technologies in nursing homes. 2022, Frontiers Media S.A. p. 957246.
17. Lussier, M., et al., Integrating an Ambient Assisted Living monitoring system into clinical decision-making in home care: An embedded case study. Gerontechnology, 2020. 19(1).
18. Hutchinson, C.S., Curbing Big Tech's IoT dominance. European Competition Journal, 2022. 18(2): p. 265-286.
19. Pridmore, J., et al., Intelligent Personal Assistants and the Intercultural Negotiations of Dataveillance in Platformed Households. Surveillance & Society, 2019. 17(1/2): p. 125-131.
20. Burdon, M. and T. Cohen, Modulation Harms and the Google Home. Surveillance & Society, 2021. 19(2): p. 154-167.
21. Harper, S., et al., Security and privacy of pet technologies: actual risks vs user perception. Frontiers in The Internet of Things, 2023: p. 1-21.

SMART DOORBELLS IN A SURVEILLANCE SOCIETY

ANJELA MIKHAYLOVA[1] [0009-0005-5731-5040], SARA WILFORD[2] [0000-0001-8562-870X], BERND STAHL[3] [0000-0002-4058-4456], LAURENCE BROOKS[4] [0000-0002-5456-8799]

Keywords: Smart doorbells, Surveillance Society, Panopticon, Ethics, Privacy.

Extended Abstract

Background

Smart doorbell technology has redefined the concept of the front door, transforming it into an intelligent surveillance hub that integrates with security systems and smart locks, syncs with voice assistants, and exchanges data with other connected devices within the smart home ecosystem [7, 13, 23]. Leading smart doorbell brands include Amazon's Ring and Blink, Google's Nest Doorbell, and Anker's Eufy. These devices are equipped with motion sensors that, when the doorbell rings or activity is detected at the property entrance, trigger high-definition cameras and microphones to begin recording audio and video, which are then stored in the cloud or on local storage. Users can remotely monitor their entryways and engage in real-time, audio-visual communication with visitors via apps.

While smart doorbells are valued for their convenience and perceived security benefits, preliminary findings from the present study's pilot survey indicate that their increasing use raises forth a spectrum of ethical issues related to surveillance, privacy, and data governance. One key insight from the pilot is the blurring of boundaries between public and private domains, particularly when surveillance practices encroach on shared spaces. As a result, smart doorbell users are empowered to surveil individuals without their knowledge or consent, thereby institutionalising private surveillance into areas such as pavements, driveways, and shared entrances [1, 12].

Furthermore, beyond the surface-level advantages of convenience and security, the pilot study suggests that smart doorbells introduce deeper questions regarding who controls access to the collected data, the extent of "consumer-initiated surveillance" [16], and the fine line between security and privacy intrusion. These preliminary insights reveal that the ethical dilemmas are not merely theoretical but are experienced by stakeholders in real-world settings. Given these complexities, this research is guided by

1 DE MONTFORT UNIVERSITY, GATEWAY HOUSE, LEICESTER, LE1 9BH, UK, ANJELA.MIKHAYLOVA@DMU.AC.UK; UNIVERSITY OF NOTTINGHAM, COMPUTER SCIENCE, NOTTINGHAM, NG8 1BB, UK, ANJELA.MIKHAYLOVA@NOTTINGHAM.AC.UK
2 DE MONTFORT UNIVERSITY, GATEWAY HOUSE, LEICESTER, LE1 9BH, UK, SARA@DMU.AC.UK
3 UNIVERSITY OF NOTTINGHAM, COMPUTER SCIENCE, NOTTINGHAM, NG8 1BB, UK, BERND.STAHL@NOTTINGHAM.AC.UK
4 UNIVERSITY OF SHEFFIELD, INFORMATION SCHOOL, THE WAVE, SHEFFIELD, S10 2AH, UK, L.BROOKS@SHEFFIELD.AC.UK

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 various stakeholders on the ethical and social implications posed by integrating smart doorbells into residential security. Drawing from the pilot survey, the research analyses the trade-offs between perceived security benefits and potential privacy intrusions. It investigates the key issue of the creeping expansion of surveillance through smart doorbells, while also addressing topics such as convenience, crime prevention, and public safety. Beyond these trade-offs, the study examines broader, unintended consequences of this technology, including its impact on social trust, community interactions, and personal autonomy. By addressing these themes, this research aims to enrich the dialogue on responsible technology use and the evolving landscape of a surveillance-driven society.

Literature Review

This literature review adopts a dual-pronged approach by drawing on both peer-reviewed academic sources to ground the discussion in established theoretical and empirical research, and non-academic literature, including legal cases, media reports, and investigative journalism, to contextualise real-world developments, public discourse, and regulatory challenges.

Despite the growing ubiquity of smart doorbells, scholarly attention to their ethical and societal implications has remained relatively limited. Additionally, much of the existing literature subsumes these devices under broader discussions of smart home ecosystems [7, 13], inadvertently overlooking the unique surveillance dynamics they introduce. However, recent academic research has begun to highlight the complex privacy challenges posed by smart doorbells [1, 13, 15, 16]. For example, Alshehri et al. [1] and Kelly [16] examine these challenges from both homeowners' and bystanders' perspectives, revealing tensions around "consumer-initiated surveillance". Moreover, Selinger and Durant [26] contend 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 [15] underscores the underacknowledged role of audio recordings in exacerbating privacy risks, noting that many users remain unaware that these devices capture sound well beyond property boundaries, a phenomenon Schleusener [27] terms a "sweeping surveillance nexus" [16, 27]. Research has further linked smart doorbell use to discriminatory surveillance practices and the disproportionate targeting of marginalised communities [17, 28].

Beyond academic discourse, legal judgements such as Fairhurst v. Woodard [12] demonstrate how unchecked residential surveillance can lead to actionable violations of privacy law. In a similar vein, Morris [21] critiques the integration of smart doorbells with law enforcement by highlighting the lack of legal safeguards and arguing that the absence of warrants for accessing footage risks undermining civil liberties. This "tech exceptionalism," as Morris [21] describes it, heightens regulatory challenges, and creates a slippery slope for pervasive surveillance and potential misuse.

This review demonstrates that the ethical challenges posed by smart doorbell technology demand deeper academic exploration and more informed regulatory ap

proaches. Yet several important questions remain unanswered regarding how these devices reshape social interactions, community dynamics, and related ethical concerns, underscoring the need for further interdisciplinary research.

Theoretical Framework

To better understand the ethical and social implications identified in the existing literature and supported by early findings form this study, this research applies two established theoretical frameworks: Foucault's concept of Panopticism and Zuboff's theory of Surveillance Capitalism. These frameworks help explain how surveillance restructures power, autonomy, and behaviour, and are used here to analyse the role of smart doorbells within contemporary surveillance society.

Watching Without Watching: The Panoptic Effect at the Doorstep

Foucault's concept of Panopticism [14] describes a surveillance model in which individuals alter their behaviour under the mere assumption of being watched, regardless of whether surveillance is active [14, 20, 25]. Smart doorbells replicate this dynamic at the domestic threshold by creating a panoptic effect – they act as digital overseers that monitor and capture any movement, be it gait, facial expressions, conversations, tone of voice, and even habitual walking routes. The constant possibility of observation can lead both residents and passersby to self-regulate their actions, suppress emotional expression, adjust daily routines, and restructure social interactions in response to perceived scrutiny.

Monetising the Front Door: Surveillance Capitalism in Practice

In Zuboff's model of Surveillance Capitalism [31], everyday interactions happening near one's home become inputs in a commodified system of surveillance. This commodification is not hypothetical – it is contractually embedded in corporate policies. Ring's Terms of Service explicitly state: "You 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 […]" [32]. Although the policy does not explicitly confirm the sale of data, it permits broad usage, including for advertising and analytics. This reflects Zuboff's concern that both users and those being surveilled are monetised and disempowered within a corporate ecosystem that commodifies domestic space and social behaviour.

Methodology

This paper draws on preliminary findings from a pilot survey conducted in a residential neighbourhood in Leicester, UK, where 30 out of 48 households had installed smart doorbells. The pilot focused on users' perspectives and early reflections revealed that these devices were beginning to influence neighbourhood social dynamics in com

plex and sometimes troubling ways. Some participants reported a growing sense of distrust towards neighbours using smart doorbells, while others described using them to gather evidence in disputes or as a deterrent against individuals with whom they had strained relationships. Participants further described instances where smart doorbell surveillance was used as a protective measure in cases of domestic violence. Others admitted to reviewing audio and video recordings of street conversations not intended for them, raising questions about unintended eavesdropping, ethical boundaries, and consent. Several users shared accounts of using the devices to gain the upper hand in neighbourhood disagreements, while others noted having shared footage with law enforcement upon request. Conversely, some participants expressed discomfort with the normalisation of surveillance, questioning whether the benefits outweigh the erosion of community trust.

Although these findings are based on a small sample of 19 households, preliminary thematic analysis revealed early patterns related to surveillance rivalry, moral ambiguity in the use of captured footage, and erosion of informal social norms and cohesion. These insights informed the design of the broader research project, which will adopt a qualitative methodology involving semi-structured interviews and focus groups in the next phase. The study will follow an interpretivist paradigm and an inductive approach, using thematic and content analysis to identify patterns, relationships, and differences across five stakeholder groups. The targeted sample will include approximately 50 to 60 participants selected through snowball and purposive sampling strategies, including primary users (i.e., smart doorbell users), non-primary users (such as neighbours, passersby, and incidental subjects), police representatives, privacy specialists, and technology experts.

Contribution

This research offers an early contribution to the underexplored area of smart doorbell surveillance by highlighting its effects on privacy, trust, and community dynamics. It enriches 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.

Acknowledgement: 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. Proceedings on Privacy Enhancing Technologies 2022(3), 99–119 (2022)
2. BBC News: Plea for doorbell footage after tree attacks, BBC News, 10 February. https://www.bbc.co.uk/news/articles/c5y7jlj9x0po, last accessed 2025/02/10
3. 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, last accessed 2025/01/05
4. 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)
5. 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, last accessed 2024/11/25
6. Bryman, A.: Social Research Methods. 4th edn. Oxford University Press, Oxford (2012)
7. Buil-Gil, D., Kemp, S., Kuenzel, S., Coventry, L., Zakhary, S., Tilley, D., Nicholson, J.: The digital harms of smart home devices: A systematic literature review. Computers in Human Behavior 145, 107770 (2023) doi:10.1016/j.chb.2023.107770.
8. 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, CA (2019)
9. Calacci, D., Shen, J.J., Pentland, A.: The Cop In Your Neighbor's Doorbell. Proceedings of the ACM on Human-Computer Interaction, pp. 1–19 (2022) doi: 10.1145/3555125
10. Dawson, C.: Practical Research Methods. How To Books Ltd., Oxford (2002)
11. Denscombe, M.: The Good Research Guide: Research Methods for Small-Scale Social Research Projects. 7th edn. McGraw-Hill Education, Maidenhead (2021)
12. Fairhurst v Woodard: Judgment. https://www.judiciary.uk/wpcontent/uploads/2022/07/Fairhurst-v-Woodard-Judgment-1.pdf, last accessed 2024/12/12
13. 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) doi: 10.1145/3411170.3411273.
14. Foucault, M.: Discipline and Punish: The Birth of the Prison. new edn., Penguin, London (1991)
15. Grauer, Y.: Video Doorbell Cameras Record Audio, Too. Consumer Reports. https://www.consumerreports.org/video-doorbells/video-doorbell-cameras-record-audio-tooa4636115889/, last accessed 2025/04/02.
16. 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) doi: 10.1007/s10603-022-09534-3
17. Kurwa, R.: Building the Digitally Gated Community: The Case of Nextdoor. Surveillance & Society 17(1/2), 111–117 (2019)
18. Maalsem, S., Sadowski, J.: The smart home on FIRE: Amplifying and accelerating domestic surveillance. Surveillance & Society 17, 118–124 (2019)
19. Magnusson, E., Marecek, J.: Doing Interview-based Qualitative Research: A Learner's Guide. Cambridge University Press, Cambridge (2015) doi: 10.1017/9781107449893
20. Manokha, I.: Surveillance, Panopticism, and Self-Discipline in the Digital Age. Surveillance & Society 16(2), 219–237 (2018)
21. Morris, J.: Surveillance by Amazon: The Warrant Requirement, Tech Exceptionalism, & Ring Security. BUJ Sci. & Tech. L. 27, 237 (2021).
22. Oates, B.J.: Researching Information Systems and Computing. SAGE Publications, London (2006)
23. Reid, L., Sisel, G.: Digital care at home: Exploring the role of smart consumer devices. Health & Place 73, 102961 (2022) doi:10.1016/j.healthplace.2022.102961
24. Sadowski, N.: Too Smart: How Digital Capitalism Is Extracting Data, Controlling Our Lives, and Taking over the World. MIT Press, New York (2020)
25. Salaudeen, M.: Panopticism and Surveillance Capitalism, Medium. https://medium.com/@mustapher07/panopticism-and-surveillance-capitalism4d6701519429, last accessed 2024/12/28

26. Selinger, E., Durant, D.: Amazon's Ring: Surveillance as a Slippery Slope Service. Science as Culture 31(4), 1–15 (2021) doi: 10.1080/09505431.2021.1983797
27. Schleusener, S.: The Surveillance Nexus: Digital Culture and the Society of Control. REAL: Yearbook of Research in English and American Literature 34(1), 175–202 (2019)
28. 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) doi:10.1145/3313831.3376255
29. 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, last accessed 2024/12/13
30. 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) doi:10.1145/3563657.3596012
31. Zuboff, S.: The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs, New York (2019)
32. Ring LLC: Ring Terms of Service. https://en-uk.ring.com/pages/terms, last accessed 2025/05/28

UNIVERSITY TEACHERS TOWARDS CHATGPT. CHALLENGES AND OPPORTUNITIES IN THE EDUCATION OF FUTURE TEACHERS

Ewa Duda[1] [0000-0003-4535-6388], Ewelina Młynarczyk-Karabin[2] [0000-0002-4684-0895], Julia Więckiewicz-Modrzewska[3] [0000-0002-6391-4665]

Keywords: ChatGPT, chatbots, higher education, academic teachers, technology enhanced education, AI ethics.

Technological advances in artificial intelligence (AI) are increasingly influencing the reality in which we live, and covering more and more areas of it. Whether we like it or not, we have to accept that it is becoming an integral part of our lives. Equally, we must accept that it is increasingly finding its way into higher education, bringing both opportunities and new challenges, including ethical ones. One of the most advanced and widely available AI tools is ChatGPT, a language model developed by OpenAI, which is widely used in various fields, including academic teaching, thanks to its ability to process and generate text. Contemporary higher education is faced with the need to adapt to the growing role of AI tools that can support teachers in their teaching, administrative and research work. At the same time, a number of ethical dilemmas arise regarding the impact of ChatGPT on academic integrity, the quality of education and the changing role of academics in the educational process.

From the perspective of academics, ChatGPT has the potential to become a valuable adjunct to the teaching process, particularly in the context of individualising teaching, increasing work efficiency and providing access to learning resources. However, its widespread use also raises a number of concerns - both in terms of student misuse of the technology and the impact on traditional teaching methods and academic values, especially in the context of the broad possibilities offered by ChatGPT [1]. As [2] point out, artificial intelligence allows for the personalisation of education by tailoring materials to the needs of individual students. ChatGPT can act as a teaching assistant, providing students with explanations at different levels, giving feedback, helping with homework, and supporting their individual development, thus complementing the educational process initiated by the academic teacher.

ChatGPT can also free academic staff from time-consuming administrative and teaching tasks. According to a study by [3], AI can assist in assessing student work, formulating tests and analysing student performance, allowing teachers to focus on more demanding aspects of teaching, such as mentoring and individual work with students.

1 Maria Grzegorzewska University, Warsaw, Poland, eduda@aps.edu.pl
2 Maria Grzegorzewska University, Warsaw, Poland, emlynarczyk@aps.edu.pl
3 Maria Grzegorzewska University, Warsaw, Poland, jwieckiewicz@aps.edu.pl

ETHICOMP2025 EXTENDED ABSTRACTS

In addition, the use of AI offers academics the opportunity to use increasingly innovative assessment methods that could not be effectively implemented using a theoretical approach alone. Furthermore, the tool can be used to quickly analyse large data sets, allowing for improved administrative processes in education [4].

The use of ChatGPT may lead to a redefinition of the role of the academic teacher in the teaching process. As [3] point out, the increasing capabilities of AI may reduce the importance of some of the teacher's functions, such as the transmission of factual knowledge, and increase the importance of skills related to mentoring, critical thinking and evaluation of sources. However, this change also risks weakening the authority of university teachers and reducing their influence on students' development, requiring new teaching strategies and ethical guidelines for the use of AI.

Empirical research on academics' attitudes towards ChatGPT reveals a mixture of favourable and unfavourable responses. On the one hand, many academics see the potential for AI to improve teaching and research, including tasks related to idea generation, assisting with literature reviews, essay writing, and improving the linguistic fluency of academic papers. On the other hand, concerns about academic integrity, the quality of AI-generated content, and the disruption of the traditional teaching model have led many academics to argue for the need for clear ethical rules governing the use of ChatGPT in academics' teaching and research [5].

[6] research suggests that AI-generated content may contain factual errors and replicate biases embedded in the training data. [7] note that the lack of transparency of AI models makes it difficult for academics to verify the accuracy and neutrality of the content provided by ChatGPT. In the context of educational ethics, this presents teachers with the challenge of having to constantly question the material generated and teach students to be critical of AI-generated content. In the context of scientific work, on the other hand, [7] emphasise that while 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 do the scientific work for the user, but provides some guidance and support.

In the article, we will present the results of a survey conducted among academic staff at one of the Pedagogical Universities in Poland. The subject of the study is academic teachers' attitudes towards the use of chatbots and tools based on large language models (LLMs) in their teaching and research work, their perspectives on the future of this technology in teaching and research work. 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 with a view to inviting them to take part 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. 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. As a result, 101 participants took part in the survey. We discarded records with less than 50% completion. Finally, we included data collected from 92 participants in the analysis, including 63 (68.5% of all participants) representing the discipline of pedagogy (34.8% of the tar-

geted 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_{age} = 43.1, SD = 9.6, range 26-73), 25 men (27.2%, M_{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 Professor's degree (3.3%). In terms of professional experience, the study sample included 21.7% of academics with less than 5 years of experience in higher education, 22.8% with 5-10 years of experience, 30.4% with 11-20 years of experience and 25.0% with more than 20 years of experience.

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 [8], 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 [8, 9].

Participants in the study indicated that ChatGPT is the tool they are most familiar with (43.5% of participants know and use it frequently, 38.0% know but use it rarely, 1.1% do not know it) and therefore use most frequently. Other chatbots were in majority unknown or used rarely (see Fig. 1), so we will focus on this aspect of the study.

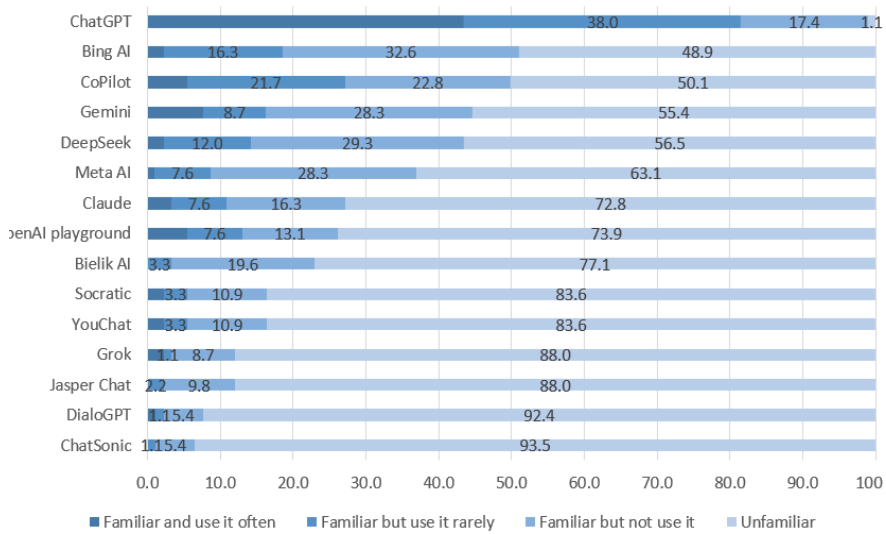


Fig. 1. Familiarity and usage of AI chatbots among academic teachers (in %), N = 92.

We then explored questions about academic teachers' stated knowledge and attitudes towards AI trainings. More than half of teachers declare they know how chatbots work. Teachers also confirm that they have read some articles or research papers on chatbots. They believe that pupils, students and teachers should participate in training on using AI. 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 ($M=3.86$, $SD=1.18$, $U = 557.0$, $p = .006$, $r = 0.29$).

Looking at the ethical issues, almost half of the respondents feel security and privacy concerns when using a chatbot. 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 ($M = 3.04$, $SD = 0.82$, $U = 119.5$, $p = .012$, $r = 0.39$).

The uniqueness of the study lies in the selection of the research sample, such as the experiences of pedagogical university staff towards ChatGPT, academics who have a direct impact on the education of future teachers. It is crucial that the development of technology goes hand in hand with ethical reflection in order to make informed and responsible use of its potential. In the full article, we will share the results of the entire study and present the attitudes of academic teachers towards the use of chatbots and tools based on large language models in their research and teaching work.

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

References

1. 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 International Journal of Management Education 21(2), 100790 (2023).
2. 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).
3. Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W.J., Heathcote, L.: The role of ChatGPT in higher education: Benefits, challenges, and future research directions. Journal of Applied Learning and Teaching 6(1), 41–56 (2023).
4. Sullivan, M., Kelly, A., McLaughlan, P.: ChatGPT in higher education: Considerations for academic integrity and student learning. Journal of Applied Learning & Teaching 6(1), 1–10 (2023).
5. Bin-Nashwan, S.A., Sadallah, M., Bouteraa, M.: Use of ChatGPT in academia: Academic integrity hangs in the balance. Technology in Society 75, 102370 (2023).
6. Karamanlıoğlu, 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).
7. Dönmez, İ., Idin, S., Gülen, S.: Conducting Academic Research with the AI Interface ChatGPT: Challenges and Opportunities. Journal of STEAM Education 6(2), 101–118 (2023).
8. 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).
9. Acosta-Enriquez, B.G., Arbulú Ballesteros, M.A., Arbulu Perez Vargas, C.G., Orellana Ulloa, M.N., Gutiérrez Ulloa, C.R., Pizarro Romero, J.M., López Roca, C.: Knowledge, attitudes, and perceived Ethics regarding the use of ChatGPT among generation Z university students. International Journal for Educational Integrity 20(1), 10 (2024).

STUDENTS' PERSPECTIVES ON CHALLENGES AND OPPORTUNITIES OF USING CHATGPT IN HIGHER EDUCATION

Ewa Duda[1] [0000-0003-4535-6388], Ewelina Młynarczyk-Karabin[2] [0000-0002-4684-0895], Julia Więckiewicz-Modrzewska[3] [0000-0002-6391-4665]

Keywords: ChatGPT, chatbots, higher education, student learning, technology enhanced education, AI ethics.

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 speech models is ChatGPT, which uses natural language processing technologies to generate text, answer questions and support teaching processes. Although its use in education is controversial and raises important ethical questions, an increasing number of students are using ChatGPT on a daily basis.

On the one hand, ChatGPT 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. 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.

The development of artificial intelligence (AI) technologies has had a significant impact on the transformation of higher education, introducing new teaching methods, personalizing the learning process and automating teaching processes. [1] highlights 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.

According to [2], ChatGPT can support students by generating texts, answering questions and helping them to 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.

1 Maria Grzegorzewska University, Warsaw, Poland, eduda@aps.edu.pl
2 Maria Grzegorzewska University, Warsaw, Poland, emlynarczyk@aps.edu.pl
3 Maria Grzegorzewska University, Warsaw, Poland, jwieckiewicz@aps.edu.pl

Despite its many benefits, the use of ChatGPT in higher education presents a number of ethical challenges. [3] highlights potential risks such as plagiarism, ghost-writing, and violations of academic integrity. The use of tools such as ChatGPT to generate academic content can lead to students using off-the-shelf solutions rather than engaging in an independent creative and analytical process. In addition, there is a risk that these algorithms may generate biased content, which can influence the formation of biases and hinder critical evaluation of the information presented. [4] highlights that over-reliance on AI tools can undermine students' ability to solve problems independently and develop their own analytical skills.

On the other hand, numerous studies point to the potential of ChatGPT as a tool that can contribute to the quality of higher education. [5] shows that ChatGPT, if properly implemented and regulated, can provide valuable support in the teaching process. Positive aspects include the ability to generate examples quickly, to help develop argumentation skills and to assist with text editing. Furthermore, the authors suggest that these technologies can help to reduce educational barriers by enabling students from marginalized groups or those with limited access to traditional forms of classroom support to access academic resources.

Empirical studies focusing on the student perspective indicate the complexity of perceptions of ChatGPT in the academic setting. On the one hand, students appreciate the benefits of quick access to information and support in the content creation process, which can increase their learning efficiency [6]. On the other hand, there are concerns about academic integrity, a decline in the quality of independently produced work and the risk of over-reliance on technology, which may undermine critical thinking [7]. Student surveys and interviews also show that the decision to use ChatGPT often depends on clear guidelines from educational institutions and transparency of policies on the ethical use of AI technology.

In this article, we present the results of a study conducted among students of a Polish pedagogical university regarding their attitudes towards the use of chatbots and tools based on large language models (LLMs) in the educational process. There are 3832 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. 445 participants took part in the survey. We discarded records with less than 50% completion. Finally, we included data collected from 408 participants (M_{age} = 22.44, SD = 4.63, range 18-51), representing 10.6% of those eligible. The sample included 373 women (91.4%, M_{age} = 22.5, SD = 4.7, range 18-51), 28 men (6.9%, M_{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).

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 [8], 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 [8, 9].

Participants in the study indicated that ChatGPT is the tool they are most familiar with (50.5% of participants know and use it frequently, 38.0% know but use it rarely, 1.2% do not know) and therefore use most frequently (other chatbots were in majority unknown or used rarely, see Fig. 1), so we will focus on this aspect of the study. Students use chatbots mainly in the educational purposes. Among these, the most common (64.6%) was clarifying complex topics from classes, lectures, read articles. Almost half of the students (48.2%) create summaries of long reading material such as articles, book chapters, handbooks. 43.5% of students declare that they generate ready answers during classes, homework. Every third student uses chatbots to make proofreading of self-written texts, to make language translation and to generate content when writing essays.

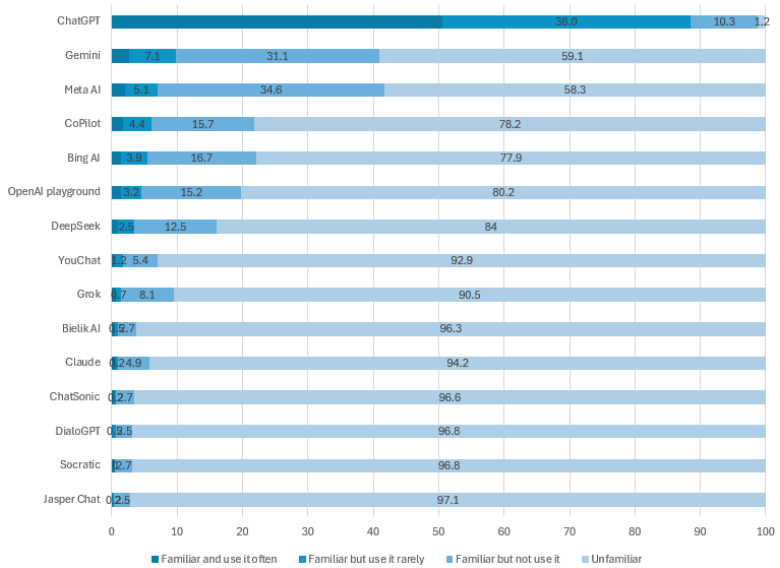


Fig. 5. Familiarity and usage of AI chatbots among students (in %). Missing answers were ignored, N = 408.

More than half of students have a positive attitude towards the use of chatbots in education (55.2%). Only one in five respondents believes that the use of chatbots should be banned in educational establishments, however they are concerned about how chatbots will impact education in the future (77.4%). Respondents have the impression that the use of chatbots is widespread among their colleagues (85.1%). Looking at the ethical issues, more than 60% of those surveyed believe that using chatbots to complete assignments and exams is cheating. Almost half of the respondents feel privacy and security concerns when using a chatbot. More than 70% of those surveyed believe that chatbots can be used to manipulate information or mislead students. Students also believe that chatbots could increase inequalities in education.

The uniqueness of the study lies in the selection of the research sample, such as the experiences, perceptions and attitudes towards ChatGPT 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. This article is an excerpt from a study that also aims to juxtapose the perspectives of students and academics from the same university on AI-based technology.

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

References

1. 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).
2. Bom, H. S. H.: Exploring the opportunities and challenges of ChatGPT in academic writing: a roundtable discussion. Nuclear Medicine and Molecular Imaging 57(4), 165–167 (2023).
3. Cordero Jr, D. A.: Addressing academic dishonesty caused by artificial intelligence tools. Journal of Public Health 46(2), e310-e310 (2024).
4. 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).
5. 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. Journal on Education 5(4), 14886–14891 (2023).
6. 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. Computers and Education: Artificial Intelligence 5, 100180 (2023).
7. Guo, Y., Lee, D.: Leveraging ChatGPT for enhancing critical thinking skills. Journal of Chemical Education 100(12), 4876–4883 (2023).
8. 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.
9. Acosta-Enriquez, B.G., Arbulú Ballesteros, M.A., Arbulu Perez Vargas, C.G., Orellana Ulloa, M.N., Gutiérrez Ulloa, C.R., Pizarro Romero, J.M., López Roca, C.: Knowledge, attitudes, and perceived ethics regarding the use of ChatGPT among Generation Z university students. International Journal for Educational Integrity 20(1), 10 (2024).

DIGITAL TRANSFORMATION OF CARE: NAVIGATING ETHICAL LANDSCAPES

Ryoko Asai[1] [0000-0002-4035-3711], Mineko Wada[2] [0000-0002-4906-0982],
Kiyoshi Murata[3] [0000-0002-5632-8078], Tatsuya Yamazaki[4] [0000-0001-8506-1731]

Keywords: Aging Society, Digital Care, Digital Technology, Ethical Issues, Elderly Care, Japan.

Introduction

Care in our Daily Lives

Care surrounds us constantly in our daily lives. Although it is imported from English language, the term care (ケア, kea) has become deeply ingrained in Japanese society. It is a word used so often that people understand the different aspects of the act without needing a specific explanation. For example, the following terms are commonly used in Japan: elder care, child care, self-care, care management, care plans and terminal care. As these examples illustrate, care is very closely tied to our lives and it is something essential to human existence, regardless of age, disability or residential place. To take an extreme example, from the moment a fertilised egg is implanted in its mother's womb, it is cared for by that mother. And when people reach old age, they receive terminal care as they approach the end of their lives. From birth to death, humans exist in a mutual exchange of care — being cared for and giving care [1].
In general, care, which is indispensable to our lives, has three main meanings [2][3].

The first is care centred on the mental or emotional act of caring for others, which can also be expressed by words such as 'consideration', 'concern' and 'thoughtfulness'. The second is care centred on actions actually performed for the sake of others, which can be expressed by phrases such as 'doing for someone' and 'looking after someone'. The third is even more specific than the second and refers to care that requires specialised and practical knowledge and skills. This third meaning of care brings to mind professional care provided by medical professionals and care workers. In any case, care can encompass not only physical care, but also emotional support, attention and concern. And we live our lives caring and being cared for, combining these three types of care in complex ways.

1 Ruhr University Bochum, Universitätsstrasse 150, 44801 Bochum, Germany, rynjyu@gmail.com
2 Hiroshima University, Minami-ku, Hiroshima 734-8553, Japan
3 Meiji University, Chiyoda-ku, Tokyo 101-8301, Japan
4 University of Toyama, Toyama-shi, Toyama 930-8555, Japan

The Care Crisis and the Promise of Digital Solutions

While care is essential for our lives, the way care is provided, especially care that requires action, is greatly affected by an increasing number of nuclear families and dual-income households, as well as the declining birthrate and escalating aging population. In other words, modern lifestyles are shifting care from an unpaid act of love within the family to an act performed by professionals in the field of care.

According to discussions on the ethics of care by Gilligan and subsequent researchers, care is essential for human beings and indispensable for their survival. However, it has traditionally been socially positioned as a role exclusively for women, and issues have been pointed out regarding the value and morality of the act of caring [4-7]. In other words, the perspective of caring for care has been socially neglected.

However, the rapid ageing of the population, combined with a declining birth rate, has highlighted the importance of care, which has traditionally been seen as a predominantly female task. In other words, there is an overwhelming shortage of human resources to care for the ever-growing elderly population. Currently, Japan is facing a serious shortage of nurses and caregivers, and is therefore accepting caregivers from abroad. However, despite its social demand and importance, care work is demanding and the salary is not high, so there is a constant need to secure personnel and human resources for care work. In particular, there is a shortage of caregivers, and the Ministry of Health, Labour and Welfare estimates that about 2.8 million caregivers will be needed by 2040, an increase of about 690,000 from 2019 [8].

Therefore, there has recently been a search for the provision of efficient and effective care services through introducing and utilising digital technology in the care sector, as seen in the development of care robots. This is the promotion of the so-called digitalisation of care. In Japan, where there are strict restrictions on accepting foreign workers, even as care workers, the introduction of digital technology in the care sector is expected to become even more active in the future. Although the Japanese government has been using the slogan “No One Left Behind” to promote user-centred digital transformation [9] and the importance of including older adults in aging and technology research and development has been increasingly emphasized in recent years [10, 11], older adults’ perspectives on a digitized society have not been explored in depth to date. Therefore, this study aims to explore how community-dwelling Japanese older adults perceive the use of digital technologies in daily life, with a particular focus on utilizing digital care. Based on the findings, we discuss the ethical issues involved in it.

Methods

This is a qualitative study guided by interpretive phenomenological inquiry [12]. Ethics approval was obtained from Ethical Committee for Epidemiology of Hiroshima University.

Participants

Eligible participants were community-dwelling older adults who were 65 years old or above and residents in Hiroshima City, Japan. To recruit participants, recruitment flyers summarizing study outlines including study objectives, eligibility criteria, data

collection protocol were circulated to Regional Community Support Centers for older adults, Older Adults' Association, and older caregivers' peer support groups. The study outlines were also announced at community-based exercise programs and salons for older adults.

Eighteen community-dwelling Japanese older adults participated in the study. Thirteen older adults were women and five were men.

Data Collection

Between August 2023 and May 2024, 10 semi-structured, in-depth interviews and two focus groups were conducted. While semi-structured, in-depth interviews were chosen as the primary method as this method enables for eliciting detailed descriptions of human experiences and perspectives [13], eight participants preferred focus groups to provide their perspectives. All interviews and focus groups were conducted in person. Both interview questions and focus group discussion questions were designed to elicit participants' experiences of adapting to digitalization in their daily lives and their related perceptions of this shift (i.e., challenges, benefits, and concerns). All the interviews and focus group discussions were digitally recorded.

Data Analysis

The digitally recorded interviews and focus group discussions were transcribed verbatim and analyzed thematically using thematic analysis techniques [14, 15]. Transcripts were descriptively coded, and initial codes were collated into broader categories, which were then amalgamated into themes.

Findings

Our analysis yielded four themes that illustrate older adults' perspectives on the use of digital technologies: 1) disapproval of dehumanization, 2) concerns about privacy and security, 3) low digital literacy, and 4) lack of access to immediate support for issues relating to digital adaptation.

Disapproval of Dehumanization

Older adults participants expressed concerns about the dehumanization associated with the current shift towards increasingly digitalized society and the negative consequences of that dehumanization. Some participants pointed out that digital technologies cannot replace human warmth that is essential to care. For example, one male older adult stated, "I think robots can never replicate human warmth" (P15M). Similarly, one female participant identified the inability of digital devices to alleviate depressed feelings:

Digital devices can't soothe a baby. Likewise, digital devices can't uplift older adults' depressed feelings, but human can, when they ask, "How are you? How are you feeling?", you know. (P14F)

Another female participant emphasized the importance of human interactions in her daily life even if they are short greetings with a casher.

I don't like using self-checkout at supermarkets and convenient stores. People just check out items by themselves and go. … I enjoy short talk with a casher, even if it's just saying "Good morning," it's great. I want that. (P2F)

Concerns about Privacy and Security

Many participants explained that they were against or hesitant of digital transformation because they were concerned about privacy and security. One participant expressed her concerns about being monitored and controlled through the digitizing of personal information. She stated, "We now have My Number (the Individual Number). I don't like it because I feel the Japanese Government monitors and tracks our personal data with it" (P8F). Another participant's account below highlighted the risks of scams and fraud associated with the use of digital technologies and her fear:

I'm worried about phone scam. I got a phone call around 8 pm the other day. A caller said, 'You'll be charged extra for your landline,' and I immediately hang up on it because I thought it must be a phone scam. (P9F)

Several participants were particularly concerned about the security of digital money transfer (e.g., online banking, credit cards, mobile payments). P16M noted:

I'm older generation…. I don't like paying by credit card especialy through the Internet. I'm okay with using a creditcard with a password in front of a casher. But I feel my credit card information can be stollen easily when I shop online and a password is not required. (P16M)

Like P16M, older adults who perceive themselves as digitally naïve may be more worried about the privacy and security issues associated with digital technologies.

Low Digital Literacy

Understanding technical terms and jargon around digital technologies was a challenge to many participants. One male participant noted, "I don't understand the meaning of the words that pop up for instruction in my computer and cell-phone. … Probably, my brain—especially on a digital part—refuses to learn and remember because it is too old" (P16M). As he associated his low digital literacy with his age, another participant below referred to her age when explaining difficulty remembering the knowledge she has gained in terms of how to use a social media and instant messaging application LINE:

The residents' association I belong to is using LINE for communication among the members. It is convenient, but I don't know how to send messages. I learned it but I forgot it. It is just my age that I easily forget how to do it when I stop using it. (P2F)

Lack of Access to Immediate Support for Issues Relating to Digital Adaptation

It was another challenge for many participants to access immediate support when they faced difficulties dealing with technological issues. One female participant

explained that she had to address her concerns about digital safety herself because she did not have resources she could rely on at hand in her family:

I installed an antivirus software in my computer because I can't ask my kids every time I wonder if it's safe to click links in emails or messages. I don't live with my kids, and my husband used to be good at computers but lately he's been saying "I forgot" when I ask him. (P6F)

When participants tried to access professional services to address technological concerns and issues, it was not easy to access them. One male participant explained:

Computer companies provide a support number, you have to wait for a long time when you call them for assistance. You'd get a message "We are experiencing a high volume of calls at this time. Please hold," and it takes at least 30 minutes until you get to talk to a support person. (P7M)

Considering that adaptation to digital society was critical for older adults, he continued to suggest that support system for them might be put in place:

Older adults who are 75 years old or above had little experience of using digital technologies at work. So if we want them to use technologies for, for example, preparing for disaster, we have to make sure that someone who knows technology very well will explain them carefully and thoroughly. (P7M)

Discussion

Perceptions and Concerns: Older Adults' Voices on Digital Technologies

According to a study by Wada et al. [9] based on interviews with Japanese community-dwelling older adults about their perceptions of the use of digital technologies in daily life, Japanese older adults have the following four concerns about care in their increasingly digitalized daily lives.

The first is a concern that care will be hampered by mechanisation or robotization, leading to dehumanisation. Dehumanization of care raises concerns about the potential loss of meaningful human interaction and also the risk of over-reliance on technology. This concern can be explained by the ongoing debate about the replacement of jobs by AI. According to Frey and Osborn [10], it is considered difficult to immediately replace humans and professions that provide care with technology or AI. Care needs to be provided in accordance with the individual circumstances of those who need it, and because care requires the consideration of life safety and security, it places not only a physical but also a psychological burden (emotional labour) on those who provide it [11].

Furthermore, it is currently difficult to replace the presence of care, where the experience of sharing a place and being with the caregiver is itself care for the cared for, with healthcare technology [12]. Kleinman further defines care as 'a personal and collective human practical experience that includes protection, practical assistance, and a sense of solidarity that is physical, emotional, interpersonal, moral, and human support' [12: p. 122]. The continued use of technology for care, which is inherently a human act, needs to be designed to play a role in human relationships through care, rather than replacing the human caregiver themselves with machines.

The second concern relates to privacy and security risks that may emerge with the adoption of digital technology. As data has become a valuable asset, it's increasingly difficult for individuals to grasp the social and economic risks associated with data breaches or misuse of personal information. These concerns of older adults are not limited to Japanese older adults, but have also been identified in research studies in Europe [13]. Furthermore, without a clear understanding of how to ensure privacy and security, these concerns make older adults reluctant to use digital technology for care. This, combined with the third concern of low digital literacy, leads to anxiety about how to use digital tools. Under the lack of digital literacy, older people easily face difficulties navigating interfaces or understanding online safety. Consequently, even in a society where digital devices are widely used, gaps in digital literacy can lead to a digital divide between people. Low digital literacy of older people can cause unexpected socio-economic losses. The problem of the digital divide in modern society needs to be understood not only as a problem of economic inequality due to digitalisation as in the past, but also as a problem that can lead to isolation of older adults and mental health problems (loneliness) in society.

This can also be caused by the fourth concern, which is the lack of access to immediate support for issues relating to digital adaptation, which makes it difficult to accept digitalisation. How should IT companies that develop and disseminate digital tools respond to the situation of older adults who are isolated in a digitalised society? These issues can be considered from the perspective of the social responsibility of digital companies and the IT industry, as well as from the perspective of the professionalism of IT professionals who design and develop digital tools.

On the other hand, these concerns can also be seen as the flip side of older adults' motivation to use digital tools and their expectations of their benefits (potential for expanding social connection, sustaining health, and avoiding using coins). It is possible to raise expectations of wanting to connect with friends, communities or society through digital tools, the desire to improve the convenience of life through digital tools, or the motivation to use them to maintain health. Introducing digitalised care into actual care practice for the sake of efficiency and convenience alone can lead to the dehumanisation of care and create an unequal power relationship between caregiver and care recipient. Mitigating these ethical and social risks requires a holistic approach to digital care that considers the interconnectedness of mind/emotion, body, skills/knowledge and connectedness, and ensures that technology serves to enhance, not diminish, the human experience of care.

Acknowledgments: This study was supported by the JSPS Fund for the Promotion of Joint International Research (International Collaborative Research) 23KK0189.

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

References

1. Asai, R.: Technohealthism: Being patients-in-waiting under the development of medical technologies. In: Arias Oliva, M. et al. (eds.) Smart Ethics in the Digital World: Proceedings of the ETHICOMP 2024, pp. 97–100. Universidad de La Rioja, Logroño (2024)

2. Kawamoto, T.: An adventure of modern ethics. Sobunsha, Tokyo (1995) (in Japanese)
3. Hiroi, Y.: Care studies: Towards transnational care. Igaku Shoin, Tokyo (2000) (in Japanese)
4. Gilligan, C.: In a different voice: Psychological theory and women's development. Harvrad University Press, Cambridge, MA. (1982)
5. Tronto, J.: Moral boundaries: A political argument for an ethics of care. Routledge, London (1993)
6. Held, V. (ed.): Justice and care: Essential readings in feminist ethics. Routledge, London (2013)
7. Okano, Y.: Care ethics. Iwanami Shoten, Tokyo (2024) (in Japanese)
8. Ministry of Health, Labour and Welfare: Measures to secure care workers, https://www.mhlw.go.jp/stf/newpage_02977.html, last accessed 2025/3/5 (2024) (in Japanese)
9. Government of Japan. Aiming for a digital society for diverse happiness. 2020. https://www.japan.go.jp/kizuna/2020/aiming_for_a_digital_society.html. Accessed 10 May 2025.
10. Boger J, Jackson P, Mulvenna M, Sixsmith J, Sixsmith A, Mihailidis A, et al. Principles for fostering the transdisciplinary development of assistive technologies. Disabil Rehabil Assist Technol. 2017;12:480–90.
11. Wada M, Grigorovich A, Kontos P, Fang ML, Sixsmith J. Addressing real-world problems through transdisciplinary working. In: Sixsmith A, Sixsmith J, Mihailidis A, Fang ML, editors.
12. Knowledge, innovation, and impact: International perspectives on social policy, administration, and practice. Springer, Cham; 2021. p. 121–9.
13. Frechette J, Bitzas V, Aubry M, Kilpatrick K, Lavoie-Tremblay M. Capturing lived experience: Methodological considerations for interpretive phenomenological inquiry. Int J Qual Methods. 2020;19.
14. Hesse-Biber SN. The practice of qualitative research. 3rd edition. Thousand Oaks, CA: Sage; 2016.
15. Braun V, Clarke V. Thematic analysis. In: Cooper H, Camic P, Long DL, Panter AT, Rindskopf D, Sher KJ, editors. APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological. Washington, DC: American Psychological Association; 2012. p. 57–71.
16. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 2006;3:77–101
17. Frey, C. B., Osborne, M. A.: The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114, 254-280 (2017)
18. Hochschild, A. R.: The managed heart: Commercialization of human feeling. University of California Press, Berkeley, CA (1983).
19. Kleinman, A.: The meaning of caring, Seishin Shobo, Tokyo (2015) (edited and translated by Kaito, A. et al.; in Japanese)
20. Klaver, N., van de Klundert, J., van den Broek, R. J. G. M., Askari, M.: Relationship between perceived risks of using mhealth applications and the intention to use them among older adults in the Netherlands: Cross-sectional study. JMIR Mhealth Uhealth, 9(8), e26845 (2021)

QUIZLY: TRANSFORMING QUIZ EXPERIENCES WITH MULTI-MODAL INPUTS FOR DIFFERENTLY ABLED USERS

FARZANA JABEEN[1], SIDRA SULTANA[2], TAHIRA ANWAR LASHARI[3], MEHVISH RASHID[4]

Keywords: Inclusive Educational Technology, Multi-modal Interaction, Accessibility for differently abled users.

Abstract

In the evolving mobile app landscape, Quizly emerges as a multimodal Android quiz application designed to enhance educational accessibility and interactivity, particularly for differently abled users. By integrating Convolutional Neural Networks (CNNs), Quizly supports hand gesture recognition, voice navigation, and gaze tracking, offering an intuitive and inclusive quiz-taking experience. Built with a robust tech stack Python for machine learning, React Native for UI, and a Node.js, MongoDB backend. The app was developed using an iterative methodology that incorporated continuous user feedback. A user study with 30 participants validated its usability and adaptability, reflecting high user satisfaction. Quizly stands out as both a technological innovation and a step forward in inclusive educational tools, delivering real-time multimodal interaction and setting the stage for future research in accessible learning platforms.

Introduction

In an era defined by rapid advancements in educational technology, ensuring accessibility and inclusivity remains a critical challenge. Despite a proliferation of digital learning tools, individuals with disabilities often encounter barriers that hinder full participation. Quizly a multimodal Android-based quiz application aims to redefine this landscape by introducing innovative interaction mechanisms tailored for differently abled users. By incorporating gaze tracking, hand gesture recognition, and voice-based navigation, Quizly empowers users with visual and motor impairments to engage with educational content on their terms. The goal is not merely technological enhancement but fostering equitable access to learning experiences. In alignment with Universal Design for Learning (UDL) principles, Quizly represents a shift towards user-centered, adaptable education platforms that serve all learners, regardless of physical ability or senso

1 NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY (NUST), ISLAMABAD PAKISTAN, FARZANA.JABEEN@SEECS.EDU.PK
2 NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY (NUST), ISLAMABAD PAKISTAN
3 NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY (NUST), ISLAMABAD PAKISTAN
4 NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY (NUST), ISLAMABAD PAKISTAN

ry limitations. This approach aligns with recommendations by Prensky [1] and Burbules et al. [2], who emphasize the need for 21st-century educational tools to support inclusivity and digital literacy. Existing quiz technologies, while often innovative in their pedagogical delivery, frequently fall short in supporting users with disabilities. Most platforms rely heavily on visual cues and touch input, excluding learners who are blind, visually impaired, or have motor impairments. Limited support for screen readers, lack of haptic feedback, and absence of alternative input modalities like voice or gaze control highlight these shortcomings [3][8][17] [24-28].

Multimodal Interaction & Technology

Quizly's primary innovation lies in its multimodal input system. The application allows users to select quiz answers through eye gaze direction, hand gestures, and voice commands. For gaze tracking, the model was trained using the ARGaze dataset—comprising over 80,000 labeled eye images at a 32x32 resolution. Facial landmark detection and pupil tracking techniques, integrated into a CNN architecture, allow the model to determine gaze direction with 83% accuracy. This enables users with limited motor function to interact simply by looking at their intended choice. For hand gestures, a deep CNN model was trained on a vast dataset of 25,000 videos and 100,000 static images at 144x144 resolution. The system supports a range of intuitive gestures like swipe left/right, tap, or hold, mapped to quiz selection actions. Additionally, Quizly includes an audio interface powered by built-in speech-to-text tools, allowing voice navigation and quiz participation for users with visual impairments. All three input modes operate in harmony, enabling redundancy and user choice key pillars in inclusive design. Previous studies have successfully implemented similar interaction models for education and assistive tech [6][12][26].

User Interface Design

The design process was rooted in empathy and inclusiveness. Leveraging Figma for UI prototyping, the team followed an iterative approach involving feedback from users with visual and motor impairments. Interface elements were carefully designed to minimize cognitive load screens were divided into six symmetrical zones to aid muscle memory, especially for blind users. Font sizes were increased for readability, and high-contrast themes were applied for low-vision users. Key navigational tools like back, menu, and search were mapped to gestures and voice commands. Haptic feedback mechanisms provide tactile cues that reinforce successful actions. Moreover, the app includes a customization menu where users can adjust the voice speed and volume of text-to-speech, change gesture sensitivity, and select input preferences. These adjustments aim to provide autonomy and foster a sense of control among users with varying needs. This aligns with Universal Design for Learning (UDL) principles advocated by Fovet [3] and Jenkins [4-5].

User Study & Evaluation

To validate the usability and accessibility of Quizly, a user study was conducted with 30 participants: 10 visually impaired, 10 with motor impairments, and 10 non-disabled control participants. Participants engaged in training sessions followed by real-time quiz interactions using their preferred input method. Performance metrics such as input accuracy, time taken, and error rate were logged. The average accuracy of interpreting gaze and gesture inputs was 89%, reflecting robust performance across modalities. Participants with disabilities reported higher engagement compared to traditional apps, attributing this to the responsive interface and non-reliance on touch input. An ANOVA test was conducted to determine if there were significant differences in usability ratings between the three groups. The test yielded an F-value of 4.500 and a P-value of 0.041, allowing us to reject the null hypothesis. This indicates that the enhanced accessibility features in Quizly significantly improved the user experience for differently abled participants. Our evaluation strategy is informed by frameworks from Bai et al. [17] and Appleton et al. [18], who explore engagement and accessibility testing methods.

Statistical Table – Usability Scores

The usability data, shown in Table 1, reveals consistently high scores across all participant groups. The control group averaged slightly higher scores, as expected due to the absence of accessibility challenges. However, the visually and motor-impaired users rated the application highly as well a strong indication that Quizly successfully mitigates typical barriers. These results support our hypothesis that multimodal input systems can level the playing field in digital education. The quantitative results support gamification theories reviewed by Alsawaier [9] and Manzano-León et al. [10].

Training Graphs

Figures 1 and 2 illustrate the model training performance for gaze tracking and hand gesture recognition, respectively. Both graphs show a significant decline in training and validation loss over epochs, confirming model convergence and learning stability. The CNN used for gaze tracking shows a smooth learning curve with minimal overfitting, while the hand gesture model displays slightly higher initial variance due to gesture complexity. Continued dataset augmentation and fine-tuning are expected to further improve accuracy. These graphs visually reinforce the reliability of Quizly's core machine learning components. These findings are consistent with learning analytics approaches discussed by Duggal et al. [26].

Conclusion & Future Work

Inclusive educational technologies are essential to ensure fairness, justice, and equal opportunity for all learners. Many existing quiz platforms exclude differently abled users by relying solely on visual and touch inputs. Quizly addresses these gaps

through multi-modal interaction, setting a new benchmark for accessibility in digital learning. The application not only proves the feasibility of CNN-driven gesture and gaze input in mobile apps but also showcases how user-centered design can deliver high usability for a diverse population. AI tools like those in Quizly must be developed with diverse, bias-free datasets to avoid disadvantaging users. Looking ahead, future versions of Quizly will incorporate real-time adaptive learning paths, voice recognition enhancements, and augmented reality features. There is also potential to integrate Quizly into larger e-learning platforms and learning management systems. As educational institutions increasingly embrace digital learning tools, solutions like Quizly will be essential in ensuring equitable access and fostering lifelong learning for all individuals. Future plans include integration of AI tutors and adaptive content, in line with emerging educational tech trends.

Table 1: Usability Scores (on a scale of 1-10)

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

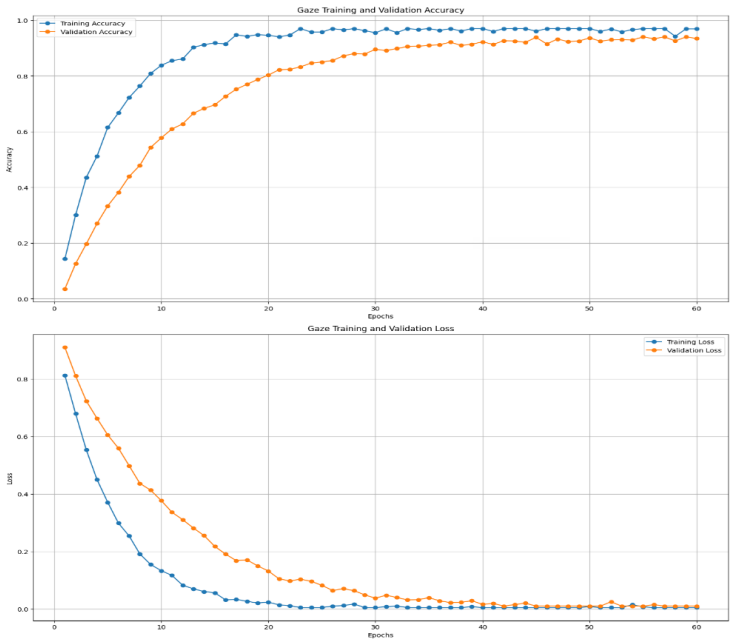


Figure 1: Gaze Tracking Model Training Curve

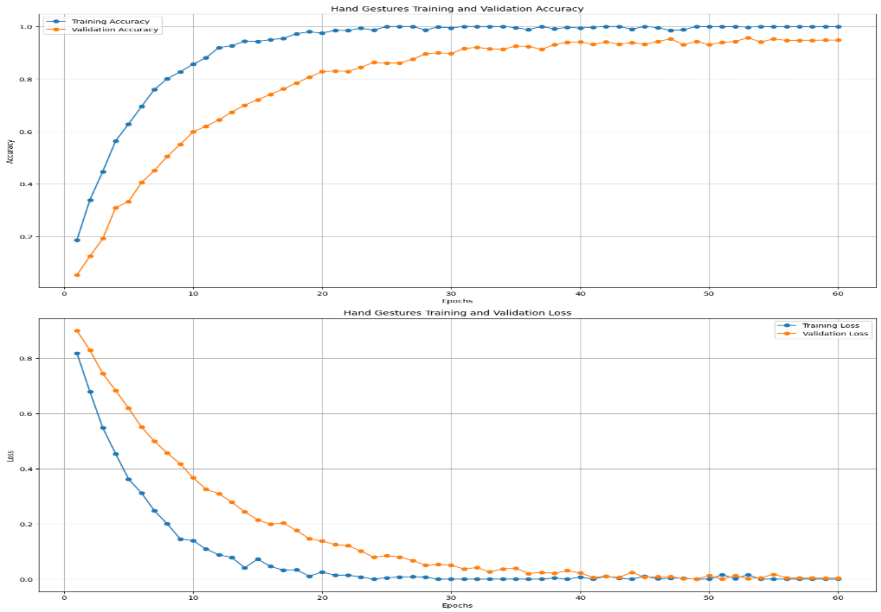


Figure 2: Hand Gesture Recognition Model Training Curve

Acknowledgments. Special thanks to our team members Zain Shafique, Muhamad Abdullah and Hamdan Ahmed.

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

References

1. M. Prensky, Education to better their world: Unleashing the power of 21st-century kids. Teachers College Press, 2016.
2. N. C. Burbules, G. Fan, and P. Repp, 'Five trends of education and technology in a sustainable future,' Geography and sustainability, vol. 1, no. 2, pp. 93–97, 2020.
3. F. Fovet, 'Universal design for learning as a tool for inclusion in the higher education classroom: Tips for the next decade of implementation,' Education journal, vol. 9, no. 6, pp. 163–172, 2020.
4. J. Ramirez, 'Educating individuals with intellectual disabilities: The US constitution, citizenship and disability,' 2021.
5. M. Jenkins, 'What is UDL?' in Unlocking Learning Potential With Universal Design in Online Learning Environments. IGI Global, 2024, pp. 1–12.
6. A. H. S. Metwally et al., 'Revealing the hotspots of educational gamification: An umbrella review,' Int. J. of Educational Research, vol. 109, p. 101832, 2021.
7. J. Parsons and L. Taylor, 'Improving student engagement,' Current issues in education, vol. 14, no. 1, 2011.
8. E. Lee and M. J. Hannafin, 'A design framework for enhancing engagement in student-centered learning: Own it, learn it, and share it,' EdTech Research & Development, vol. 64, pp. 707–734, 2016.
9. R. S. Alsawaier, 'The effect of gamification on motivation and engagement,' Int. J. of Information and Learning Technology, vol. 35, no. 1, pp. 56–79, 2018.

10. A. Manzano-León et al., 'Between level up and game over: A systematic literature review of gamification in education,' Sustainability, vol. 13, no. 4, p. 2247, 2021.
11. M. Jones et al., 'Science to practice: Does gamification enhance intrinsic motivation?' Active Learning in Higher Education, vol. 24, no. 3, pp. 273–289, 2023.
12. E. Sanchez et al., 'Gamification,' in Encyclopedia of Education and Information Technologies. Springer, 2020, pp. 816–827.
13. Z. Luo, 'Gamification for educational purposes: What are the factors contributing to varied effectiveness?' Education and Information Technologies, vol. 27, no. 1, pp. 891–915, 2022.
14. M. Alahmari et al., 'Trends and gaps in empirical research on gamification in science education,' Contemporary Educational Technology, vol. 15, no. 3, 2023.
15. W. Oliveira et al., 'Tailored gamification in education: A literature review and future agenda,' Education and Information Technologies, vol. 28, no. 1, pp. 373–406, 2023.
16. R. Huang et al., 'The impact of gamification in educational settings on student learning outcomes: A meta-analysis,' EdTech Research and Development, vol. 68, pp. 1875–1901, 2020.
17. S. Bai et al., 'Does gamification improve student learning outcome?' Educational Research Review, vol. 30, p. 100322, 2020.
18. E. F. Barkley and C. H. Major, Student engagement techniques: A handbook for college faculty. John Wiley & Sons, 2020.
19. A. N. Saleem et al., 'Gamification applications in e-learning: A literature review,' Technology, Knowledge and Learning, vol. 27, no. 1, pp. 139–159, 2022.
20. A. Aibar-Almazán et al., 'Gamification in the classroom: Kahoot! as a tool for university teaching innovation,' Frontiers in Psychology, vol. 15, p. 1370084, 2024.
21. G. L. Chan et al., 'Gamification as technology enabler in SEN and DHH education,' Education and Information Technologies, vol. 27, no. 7, pp. 9031–9064, 2022.
22. V. Arufe Giráldez et al., 'Can gamification influence the academic performance of students?' Sustainability, vol. 14, no. 9, p. 5115, 2022.
23. Z. Zainuddin et al., 'The role of gamified e-quizzes on student learning and engagement,' Computers & Education, vol. 145, p. 103729, 2020.
24. H. Moreira and M. L. L. Freire, 'Promoting formative assessment with Quizizz,' Ciencia Latina Revista Científica Multidisciplinar, vol. 8, no. 2, pp. 590–604, 2024.
25. A. F. Azizah et al., 'Improving the interaction of autistic children through eye tracking using gamification design framework,' in COSITE 2021, IEEE, pp. 57–62.
26. K. Duggal et al., 'Gamification and machine learning inspired approach for classroom engagement and learning,' Mathematical Problems in Engineering, vol. 2021, p. 9922775.
27. K. S. Narayanan and A. Kumaravel, 'Hybrid gamification and AI tutoring framework using ML and ANFIS,' J. of Advanced Research in Applied Sciences and Engineering Tech., vol. 42, no. 2, pp. 221–233, 2024.

BIOMETRICS AS BORDER CONTROL

Marie Eneman[1], Jan Ljungberg[2]

Introduction

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 risk/fear of terrorism and crime have redefined the function of borders within the Schengen area under political pressure to increase security. Simultaneously, the intensified surveillance has been criticised for risks for fundamental democratic values, particularly 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-institutional infrastructure, that will be implemented simultaneously across all 29 Schengen states. There have been repeated delays due to the system's scale and technical and legal complexity, now planned to be implemented in October 2025. 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.

Against this background, the aim of this paper was to explore ethical 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 supra-institutional infrastructure.

Supra-institutional 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, 2023; Trauttmansdorff, 2024). Scholars (Amoore, 2021, 2024a, 2024b; Amon, 2022; Mau, 2023) 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 (2023) 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.

1 Department of Applied IT, University of Gothenburg, Sweden
2 Department of Applied IT, University of Gothenburg, Sweden

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 (Haggerty & Ericson, 2000). EES is an illustrative example how surveillance operates through a surveillant assemblage and here operational images (Farocki, 2004) play a central role. In this infrastructure, border control is no longer purely the domain of officials exercising discretionary judgement but rather becomes a distributed process. Although the biometric deep border (Amoore, 2021) has become an almost ubiquitous feature in the fight against terrorism and crime, it also encompasses ambiguous and conflictual elements which makes it questionable.

Research Design

The empirical material for this study was generated through field visits, interviews, and document analysis (Alvesson & Deetz, 2021). We have conducted four airport visits, 20 interviews and document analysis. The following section presents the identified themes.

Ethical Dilemmas in Airport Biometric Border Control

The airport border as a material and institutional assemblage Airports are shaped by interwoven physical, digital, and institutional structures. In Sweden, the state-owned company Swedavia manages airport operations, infrastructure, and security procurement, while parts of airport security are delegated to Securitas. The police oversee border control and immigration, including the content of passport kiosks owned by Swedavia. Both Swedavia and the police operate separate surveillance systems. Customs handle goods control. Collaboration between these actors is shaped by legal frameworks, digital system complexity, and institutional boundaries.

Efficiency in passenger flows versus security Automated border control aims to manage rising numbers of non-Schengen travellers without expanding staffing. In Sweden, full automation is not planned; instead, biometric systems will assist border guards in identity verification. This enhances detection of identity fraud, such as look-alike cases, but increases workload, as suspected cases require further processing. Although facial images can be manipulated, such attempts remain rare. Security improves, but not flow efficiency, as the system depends on both technology and human judgement.

Increased Documentation

The adoption of biometric technologies is expected to result in more extensive documentation at border controls, particularly as the use of so-called enhanced controls becomes more frequent. Enhanced control involves escorting the individual to a separate area where two border officers conduct detailed questioning regarding identity and the

purpose of entry into Sweden. Travel documents are also subjected to close inspection, including microscopic examination.

Technical and Infrastructure Challenges

These challenges concern both the establishment of a robust infrastructure and the integration of new biometric systems with existing frameworks to achieve interoperability. They are also context-specific, depending on how particular technologies function in operational settings. For instance, the physical placement of facial recognition cameras during testing has posed difficulties, requiring manual adjustment to align with passengers' varying heights.

Infrastructural Asymmetries and Organisational Frictions

Readiness and willingness to implement the EES vary considerably across Schengen member states. High-traffic airports often face greater logistical complexity in upgrading systems without causing significant delays. In the Swedish context, postponements have also been attributed to the expectation that more advanced technology will become available, leading to a preference for deferral.

Legal and Regulatory Barriers

Swedish legislation required revision to permit the collection and processing of biometric data such as fingerprints and facial images. Under the amended legal framework, the Swedish Police Authority is responsible for facilitating access to the EES for other relevant authorities and managing personal data processing within the system. Further clarification has been provided regarding the relationship between EES regulations and broader data protection law, including the conditions under which foreign nationals are obliged to provide biometric identifiers for verification.

Detection and Prevention of Serious Crime and Terrorism

Law enforcement authorities and Europol may access EES data for the investigation and prevention of terrorism and other serious crimes, provided the request is proportionate and justified by overriding public security concerns. When biometric data collected through the EES is processed for law enforcement purposes, it falls under the scope of Directive (EU) 2016/680, which permits certain exceptions from individual rights, including access, rectification, and erasure. This legal framework creates a significant ethical tension between public security and individual data rights especially the right to erasure, or the Right to be Forgotten (GDPR).

Privacy, Sorting and Profiling

Border control practices involve the profiling of travellers, including the observation of behavioural cues while queuing or during interactions with border officials.

Such profiling is closely tied to broader processes of risk assessment, wherein border authorities evaluate potential risks related to irregular migration, threats to public order or security, human trafficking, and the unauthorised movement of minors e.g. cases of suspected child abduction. Importantly, the selection of individuals for additional control must adhere to non-discrimination principles. orientation.

Travellers' Right to Privacy

Travellers occasionally express reluctance to have their biometric data collected, often in relation to being photographed, at border control points. This often prompts explanatory dialogue in which officers are required to clarify the legal basis for biometric registration. Border crossings can be experienced as high-pressure situations, which may affect individuals' capacity to fully comprehend the implications of biometric data collection. findings also underscore the importance of upholding human dignity during the biometric enrolment process and ensuring that it is carried out in a non-discriminatory manner.

Conclusions

While biometric technologies hold the potential to increase efficiency and security, they simultaneously give rise to significant ethical and legal dilemmas especially concerning democratic values and rights such as freedom of movement and the right to privacy. In this context, the right to erasure under the GDPR emerges as particularly prominent in relation to biometrics. As biometric borders continue to evolve, there is a need for an ongoing critical and ethical discussion, not only in relation to the technology but also in relation to the power dynamics and normative assumptions they create and reinforce.

References

1. Amoore, L (2024a) Bordering worlds, modelling borders. Political Geography, 109, 103016.
2. Amoore, L (2024b) The deep border. Political Geography, 109.
3. Amoore, L., & Raley, R (2017) Securing with algorithms: Knowledge, decision, sovereignty. Security Dialogue, 48(1), 3–10
4. Ball, K & Webster, W (2020) Surveillance and Democracy in Europe, Routledge
5. Donnan, H., & Wilson, T. M. (1999) Borders: frontiers of identity, nation and state. Berg
6. European Commission, COM(2016) 205 final
7. Farocki, H (2004) Phantom Images, Public, (29), pp. 12-22.
8. Guild, E (2021) Schengen Borders and Multiple National States of Emergency: From Refugees to Terrorism to COVID-19. European Journal of Migration and Law, 23(4), 385-404.
9. Guild, E., Brouwer, E., Groenendijk, K., & Carrera, S (2016) What is happening to the Schengen borders?, Liberty and Security in Europe, No. 86.
10. Jeandesboz, J (2016) Border security and the politics of risk: The EU and the securitisation of irregular migration. In G. Lazaridis, C. Campani, & A. Benveniste (Eds.), The securitisation of migration in the EU, pp. 115–136, Palgrave Macmillan.
11. Johnson, C., Jones, R., Paasi, A., Amoore, L., Mountz, A., Salter, M., & Rumford, C (2011) Interventions on rethinking 'the border' in border studies. Political Geography, 30(2), 61–69.

12. Khan, N., & Efthymiou, M (2021) The use of biometric technology at airports: The case of customs and border protection (CBP). International Journal of Information Management Data Insights, 1(2).
13. Lyon, D (2003) Surveillance as social sorting: privacy, risk, and digital discrimination. Routledge.
14. Lyon, D (2009) Identifying Citizens: ID Cards as Surveillance, Polity Press.
15. Mau, S (2022) Sorting Machines, John Wiley and Sons LTD.
16. Simmel, G (1997) The sociology of space, In D. Frisby & M. Featherstone (Eds.), Simmel on culture: Selected writings, pp. 137–170, Sage.
17. Trauttmansdorff, P (2024) The Digital Transformation of the European Border Regime: The Power and Perils of Imagining Future Borders, Bristol University Press.

SUPPORTING INDEPENDENT LIVING FOR OLDER ADULTS: THE ROLE OF DIGITAL TECHNOLOGY AND AI IN SWEDEN AND JAPAN

Ryoko Asai[1] [0000-0002-4035-3711], Iordanis Kavathatzopoulos[2] [0000-0003-3806-5216], Kiyoshi Murata[3] [0000-0002-5632-8078], Mineko Wada[4] [0000-0002-4906-0982]

Keywords: Artificial Intelligence, Aged Care, Digital Technology, Ethical Issues, Home Care.

Extended Abstract

Demographic Change and Independent Living

Nowadays, we globally face demographic change caused by aging populations. Japan and Sweden, the two countries examined in this study, are no exception. Life expectancy in both is among the highest in the world: Approximately 83 years old in Sweden and 84 years old in Japan [1]. Demographic aging and other structural changes will increase social expenditures, such as medical costs and public pension spending [2].

By 2040, around 25 percent of the Swedish population will be 65 or older, and many in this age group will be active and healthy for a considerable period of their older years [3] Japan faces a similar demographic trend; it is estimated that by 2040, the population aged 65 and over will constitute approximately 35 percent of the total population [4]. As the population ages and social security costs rise, supporting independent living becomes a pressing issue in the care sector. For older people, being healthy and living independently is essential for maintaining a sufficient quality of life and establishing a sustainable and participatory relationship with communities and society. However, physical weakness and declining cognitive abilities can make it difficult to live a completely independent life. Therefore, support for independent living enables older people to maximise their independence.

Support for independent living among older adults encompasses a range of services and assistance, extending beyond formal care services provided by public institutions to include informal support from family members, irrespective of co-residence. Publicly available formal services encompass home care services and home medical care. Home care services, designed to enable older adults to maximise independent

1 Ruhr University Bochum, Universitätsstrasse 150, 44801 Bochum, Germany, rynjyu@gmail.com
2 Uppsala University, Lägerhyddsvägen 1, 752 37 Uppsala, Sweden
3 Meiji University, Chiyoda-ku, Tokyo 101-8301, Japan
4 Hiroshima University, Minami-ku, Hiroshima 734-8553, Japan

living within their homes, involve visits from home care workers (referred to as "home helpers" in Japan and "Hemtjänst" in Sweden). These professionals provide a range of support, including assistance with activities of daily living (ADL) such as eating, excretion, and bathing (physical care), as well as instrumental activities of daily living (IADL) such as cleaning, laundry, shopping, and cooking (lifestyle assistance). Furthermore, specialised care businesses offer services such as transportation assistance to medical appointments, including support with boarding, transferring, and alighting from vehicles. Home care services are tailored to individual needs, with the content and scope of services varying based on the degree of required assistance. A collaborative process involving the care recipient, their family, and a care manager determines the specific care needs, culminating in the development of a personalised care plan. The governments of both countries are prioritising policies that promote independent living among older adults to reduce social welfare costs and address the shortage of elderly care services.

The Role of Technology in Independent Living for Older Adults

Digital technology and artificial intelligence (AI) are revolutionising elderly home care, empowering aging individuals to maintain independence while enhancing safety and well-being. Wearable devices and health-monitoring tools track vital signs in real-time, enabling early detection of health issues and reducing the need for frequent doctor visits. AI-powered virtual assistants and chatbots provide personalised care, from medication reminders and exercise encouragement to combating loneliness through companionship. Robots and AI-driven devices assist with daily tasks, including mobility support, meal preparation, and home safety monitoring.

Smart home technology, exemplified by companies like Careium in Sweden, uses internet-connected devices to monitor seniors' environments and alert caregivers to potential problems, such as falls or prolonged inactivity [5]. Telemedicine is another significant advancement, allowing seniors to have video consultations with healthcare professionals, reducing the need for in-person visits [6][7]. Telemedicine platforms, like Min Doktor and Livi, allow seniors to consult remotely with healthcare professionals, making care access easier [8][9]. AI-powered fall detection devices (e.g., Safe Hub, Doro) automatically alert caregivers or emergency services, while robotic assistants like ElliQ offer companionship, medication reminders, and encourage physical activity [10]. AI-driven predictive health monitoring tools (e.g., Pilloxa) help ensure medication adherence and predict potential health risks by analysing individual data patterns [11]. Cognitive health apps (e.g., MindMate) support individuals with dementia through memory exercises and daily reminders [12]. Virtual reality platforms are also utilised to combat social isolation, offering engaging experiences like virtual travel and social interaction.

While Sweden's approach focuses on direct benefits for seniors, Japan emphasises supporting caregivers. For example, Panasonic's digital care management service uses IoT monitoring to provide valuable data insights for care plan development, while NEC offers remote training for care professionals to address staffing shortages and improve care quality. These digital tools and AI technologies hold immense potential to

transform elderly home care, promoting independence, enhancing safety and well-being, and supporting family caregivers, leading to more sustainable and compassionate care solutions for aging populations.

The Ethical and Practical Considerations of Digital Technology and AI in Aged Care at Home

While IT tools and AI offer significant benefits for elderly care at home, there are several risks, medical as well as ethical, that need to be considered. One major concern is privacy and data security. As wearable devices, sensors, and AI-driven solutions collect sensitive health data, there is the potential for data breaches or unauthorised access, putting personal information at risk. Ensuring robust encryption and secure storage of this data is critical. Another risk involves the reliability and accuracy of the technology. AI systems and monitoring devices may malfunction or provide incorrect data, leading to false alarms or missed health issues. This could result in unnecessary stress for caregivers or, conversely, delay timely intervention for the elderly.

Additionally, technology may not be fully accessible to all seniors, especially those with limited digital literacy or cognitive impairments, potentially exacerbating existing inequalities in access to care. Over-reliance on technology could reduce the human element of caregiving. While AI can offer assistance, it cannot replace the emotional connection and personalised care that human caregivers provide. Balancing technology with human involvement is essential to ensure that elderly individuals receive comprehensive, compassionate care.

It is therefore important to consider how technology can be leveraged to provide family members with the skills and support they need as caregivers. Many dedicated family caregivers lack formal training in medical or caregiving tasks. The integration of IT and AI tools can offer substantial support. These tools provide both training resources and real-time assistance to enhance caregiving.

Digital platforms and mobile applications offer self-paced training modules that teach family caregivers how to handle basic medical tasks like administering medication, checking vital signs, or using medical devices. These tools provide step-by-step instructions and videos to improve caregiver competence. Additionally, AI-powered virtual assistants and chatbots can guide family members in daily caregiving routines, providing reminders for medications, doctor's appointments, or physical exercises, ensuring that important tasks are not forgotten. Remote monitoring systems, such as wearable devices or in-home sensors, help caregivers keep track of the elderly person's health metrics in real time, such as heart rate, activity levels, or sleep patterns. AI can analyse this data to detect potential health risks, like falls or abnormal vital signs, and alert caregivers immediately. This proactive approach enables early intervention and reduces the stress of constant surveillance. Telemedicine and virtual consultations provide caregivers with direct access to healthcare professionals for advice and guidance without needing to leave home. Family members can discuss concerns, clarify medical instructions, or receive expert opinions, ensuring proper care.

Emotional support is also vital, and AI tools can assist here by offering stress-relief strategies, such as guided meditation programs, or even companionship through virtual pets or social interaction platforms for both caregivers and the elderly. By combining training, monitoring, and support, IT and AI tools may empower family members to deliver safe, effective care, reducing the burden and enhancing the quality of life for everyone involved.

Digital technology and AI hold immense potential to transform elderly care, promoting independent living, improving safety and well-being, and supporting family caregivers. However, realising this potential requires careful attention to the ethical and practical considerations discussed, including privacy concerns, data security, technological reliability, and equitable access. A balanced approach that integrates technology with human-centred care, prioritising the emotional and social needs of older adults, is essential for creating sustainable and compassionate solutions. Future research can work on developing best practices for ethical implementation and ensuring that technology empowers both older adults and their caregivers.

Acknowledgments: This study was supported by the JSPS Fund for the Promotion of Joint International Research (International Collaborative Research) 23KK0189.

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

References

1. The World Bank: Life expectancy (data), https://data.worldbank.org/indicator/SP.DYN.LE00.IN?view=map&year=2019, last accessed 2025/3/31 (2021)
2. OECD: Social spending (indicator). https://doi.org/10.1787/7497563b-en (2021)
3. Swedish Institute: Elderly care in Sweden: Sweden's elderly care system aims to help people live independent lives, https://sweden.se/life/society/elderly-care-in-sweden, last accessed 2025/3/31 (2021)
4. Ministry of Health, Labour and Welfare: About the population of Japan, https://www.mhlw.go.jp/stf/newpage_21481.html, last accessed 2025/3/31 (n.d.) (in Japanese)
5. Careium: Framtidens vård är här, https://www.careium.com/sv-se/om-careium/future-of-care/, last accessed 2025/3/31 (n.d.)
6. Hansson, J., Blomqvist Carlsson, N.: Äldre patienters upplevelse av digital rådgivning inom primärvården. Malmö universitetet, Hälsa och samhälle (2024)
7. Mester, F.: Äldres uppfattningar av digitala tjänster. Mälardalens universitet, Akademin för hälsa, vård och välfärd (2024)
8. Min Doktor: https://www.mindoktor.se, last accessed 2025/3/31 (n.d.)
9. Livis Omsorg: https://livissomsorg.se, last accessed 2025/3/31 (n.d.)
10. ElliQ: Your AI sidekick for happier, healthier aging, https://elliq.com/?srsltid=AfmBOoouLYCdcnYv_UjdZwWHJx8La5h9EU7RZUuhEdTBPbJ4X_dr0uA, last accessed 2025/3/31 (n.d.)
11. Pilloxa: https://pilloxa.com, last accessed 2025/3/31 (n.d.)
12. Mindmate: https://www.mindmate-app.com, last accessed 2025/3/31 (n.d.)
13. Panasonic: Digital care management, https://tech.panasonic.com/jp/lifelens/dcm.html, last accessed 2025/3/31 (n.d.)
14. NEC: Remote functional training support services, https://jpn.nec.com/rtrepo/index.html, last accessed 2025/3/31 (n.d)

INTEGRATED METHODS OF PERSONAL DATA PROTECTION IN THE CONTEXT OF MODERN CYBER THREATS

Sabina Szymoniak[1] [0000-0003-1148-5691], Mariusz Kubanek[2] [0000-0001-9651-9525]

Keywords: Personal data protection · Encryption · Anonymisation · Cybersecurity.

Extended Abstract

Among all the resources available in modern society, personal data are among the most precious. Organisations that collect and process it in an unprecedented way risk exploiting it. Some events are unauthorised access, data breaches, or data exploitation for business purposes against user consent. At the same time, users are becoming increasingly aware of the importance of protecting privacy, which forces organisations to implement adequate protection mechanisms [5].

The key threats to personal data include cyberattacks, the problem of improper data management, and Big Data practices and profiling. One of the most serious threats to the personal data of network users is cyberattacks. Cybercriminals use advanced techniques, including social engineering, to gain access to personal data. Hence, the most dangerous data attacks include phishing [16], ransomware [7], and DDoS [6] attacks. Phishing is a social engineering technique to obtain confidential information by impersonating trusted institutions, such as passwords, bank details, or credit card numbers. These attacks are often carried out using fake emails or websites that look almost identical to the original. Ransomware is malware that encrypts the victim's data and demands a ransom to unlock it. In the latest attack, the so-called double extortion, attackers first steal data and then threaten to reveal it if payment is not made. DDoS (Distributed Denial of Service) attacks overload server or network resources by sending many requests from infected devices simultaneously. Their goal is to prevent legitimate users from accessing online services, which can lead to serious financial and reputational losses [12]. Figure 1 shows the key threats to personal data.

An improperly conducted data management process is equally dangerous for personal data. Organisations often store data inconsistent with best practices, such as without encryption or in outdated systems. Such negligence leads to inadvertent leaks that can have serious legal and reputational consequences. Furthermore, collecting large amounts of data for analytical purposes creates a risk of privacy breaches. Profiling algorithms used in marketing and politics can lead to discrimination or manipulation [2].

1 Department of Computer Science, Czestochowa University of Technology, Czestochowa, Dabrowskiego 69, Poland, sabina.szymoniak@icis.pcz.pl

2 Department of Computer Science, Czestochowa University of Technology, Czestochowa, Dabrowskiego 69, Poland, mariusz.kubanek@icis.pcz.pl

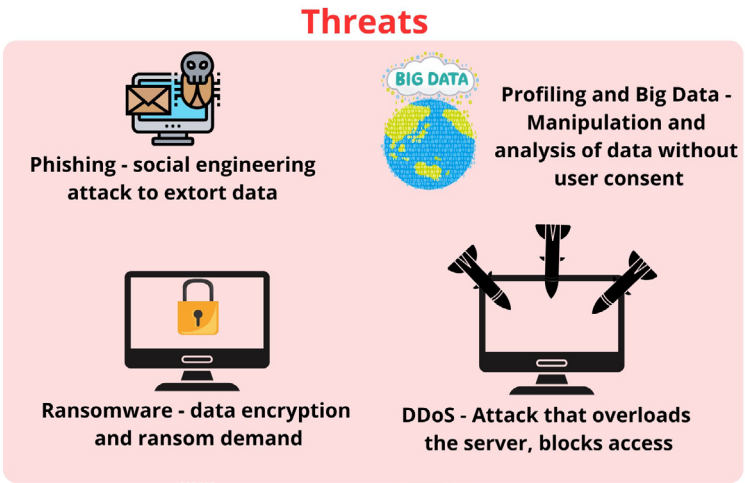


Fig.1. Key threats to personal data.

An essential element aimed at improving the security of personal data is the implementation of appropriate mechanisms for protecting personal data. Here, we can point out the General Data Protection Regulation (GDPR) [3]. The GDPR introduced uniform data protection principles in the European Union. In addition, it imposed obligations on data controllers, such as reporting breaches within 72 hours, ensuring the right to be forgotten, and data protection impact assessments. The NIS2 directive [13] (Network and Information Security Directive 2) is also worth mentioning. NIS2 is the updated NIS Directive (2016/1148) adopted by the European Union to improve cybersecurity within the Member States. Thus, the organisation has to adapt all security procedures, identify critical IT resources, implement protective mechanisms, and notify all incidents to the appropriate authorities [15].

Equally important are technologies that support data protection. Encryption is the basic mechanism for protecting against unauthorized access. Data anonymization eliminates the possibility of linking data to a specific person, while pseudonymization reduces the risk of identification and enables its processing for analytical purposes. From a network perspective, threat detection and prevention systems are essential. These systems monitor network traffic, identify potential threats, and automatically respond to incidents. They use machine learning algorithms to analyse attack patterns. Encryption algorithms AES-256 [8] and RSA [11] protect data in rest and transit. Measures such as k-anonymity or differential privacy make anonymization feasible, reducing the possibility of users' identification. IDS/IPS tools such as Snort [1] or Suricata [9]detect threats. Together, these technologies provide end-to-end protection from data leakage and cyberattacks. Nevertheless, user awareness should always be addressed. It is a key element of data protection. Educational campaigns, training, and certificates in information security helps minimize the risk resulting from human errors [4].

Personal data protection is not merely a matter of legislation but, above all, responsibility for users' privacy on the part of companies. Businesses have to be honest,

considering the users' will and not playing tricks on customers with tricks such as mass profiling. Ethics regarding new technology is also essential business objectives only with artificial intelligence or Big Data can't be solved at the expense of privacy. In reality, it's not simply compliance with legal rules but thoughtful decision-making: is the information indeed required? Is using it harmful to users? Such a framework instils confidence and facilitates building technologies that add to privacy rather than diminish it [10].

As new technology emerges, so do new ways of compromising privacy. Augmented reality, quantum computers, and better artificial intelligence models can change data collection, processing, and protection. Perhaps the most immediate issue could be AI's growing capability to predict and profile people using increasingly granular behavioural information. In the future, new legal laws and more efficient protection means will have to be created and enforced to maintain the speed of technological development.

This article will focus on the future of personal data protection. Here, we can point out two leading technologies that will significantly influence changes in this context. First, there will be issues related to artificial intelligence. Artificial intelligence protects personal data, for example, by automatically detecting anomalies. However, its development also creates new challenges, such as the possibility of creating advanced methods of breaking security. Blockchain technologies can provide greater transparency and data processing security thanks to ledgers' immutability and distributed nature. In addition, we should also expect that in the future, legal regulations will have to consider new technologies, such as the Internet of Things (IoT) and quantum computing [14].

Personal data protection requires an integrated approach that combines legal regulations, modern technologies and user education. Despite the growing challenges, the development of protection mechanisms and increased social awareness allow for adequate privacy protection. Adapting protection strategies to new threats and opportunities will be crucial in an era of dynamic technological changes.

References

1. Caswell, B., Foster, J.C., Russell, R., Beale, J., Posluns, J.: Snort 2.0 intrusiondetection. Syngress Publishing (2003)
2. Fabrègue, B.F., Bogoni, A.: Privacy and security concerns in the smart city. SmartCities 6(1), 586–613 (2023)
3. 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 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)
4. 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)
5. Javaid, M., Haleem, A., Singh, R.P., Suman, R.: Towards insighting cybersecurityfor healthcare domains: A comprehensive review of recent practices and trends. Cyber Security and Applications 1, 100016 (2023)
6. 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. Computer Science Review 53, 100661 (2024)
7. McIntosh, T., Susnjak, T., Liu, T., Xu, D., Watters, P., Liu, D., Hao, Y., Ng,A., Halgamuge, M.: Ransomware reloaded: Re-examining its trend, research and mitigation in the era of data exfiltration. ACM Computing Surveys 57(1), 1–40 (2024)

8. Mohammed, Z.K., Mohammed, M.A., Abdulkareem, K.H., Zebari, D.A., Lakhan,A., Marhoon, H.A., Nedoma, J., Martinek, R.: A metaverse framework for iot-based remote patient monitoring and virtual consultations using aes-256 encryption. Applied Soft Computing 158, 111588 (2024)
9. Moraleda-Berral, P., Gálvez, R., Martínez-Nevado, E., Pérez de Quadros, L., García, J., de la Riva-Fraga, M., Barrera, J.P., Estévez-Sánchez, E., Cano, L., Checa, R., et al.: First clinical cases of leishmaniosis in meerkats (suricata suricatta) housed in wildlife parks in madrid, spain. Parasites & Vectors 18(1), 1–11 (2025)
10. Papadimitriou, S., Virvou, M.: General data protection regulation and adaptiveeducational games. In: Artificial Intelligence—Based Games as Novel Holistic Educational Environments to Teach 21st Century Skills, pp. 253–275. Springer (2025)
11. Rivest, D.R., Shamir, A., Adleman, L.: Rsa (cryptosystem). Arithmetic AlgorithmsAnd Applications p. 19 (1978)
12. Saeed, S., Altamimi, S.A., Alkayyal, N.A., Alshehri, E., Alabbad, D.A.: Digitaltransformation and cybersecurity challenges for businesses resilience: Issues and recommendations. Sensors 23(15), 6666 (2023)
13. Schmitz-Berndt, S.: Defining the reporting threshold for a cybersecurity incidentunder the nis directive and the nis 2 directive. Journal of Cybersecurity 9(1), tyad009 (2023)
14. Szymoniak, S., Depta, F., Karbowiak, Ł., Kubanek, M.: Trustworthy artificial intelligence methods for users' physical and environmental security: A comprehensive review. Applied Sciences 13(21), 12068 (2023)
15. Szymoniak, S., Siedlecka-Lamch, O., Kurkowski, M.: Sat-based verification of nspkprotocol including delays in the network. In: 2017 IEEE 14th International Scientific Conference on Informatics. pp. 388–393. IEEE (2017)
16. Varshney, G., Kumawat, R., Varadharajan, V., Tupakula, U., Gupta, C.: Antiphishing: A comprehensive perspective. Expert Systems with Applications 238, 122199 (2024)

DEEPFAKE AND ETHICS – ANALYSIS OF THE RISKS OF IMAGE AND TEXT MANIPULATION

MARIUSZ KUBANEK[1] [0000-0001-9651-9525], SABINA SZYMONIAK[2] [0000-0003-1148-5691]

Keywords: deepfake · ethics · disinformation · image and text manipulation,

Extended Abstract

Developing deepfake technology, i.e., generating realistic images, video, and text using artificial intelligence (AI) [14], has opened up new entertainment, education, and business possibilities. However, this technology carries serious ethical and social threats. Analysing the threats resulting from deepfake technology, several important domains of the modern world can be identified that can be affected by image and text manipulation. Also, many other methods for collecting and performing data support deepfake technology [12,13]. Figure 1 summarises domains where deepfake technology can be used.

The first domain worth mentioning is disinformation and social destabilisation. Deepfake technology can create video or audio materials that seem authentic but mislead recipients. An example would be a politician's fake speech, which aims to cause social chaos, manipulate information, and strengthen propaganda [11].

The second domain of the modern world, where deepfakes can play a key role, is cyberbullying and privacy violations. An example here is deepfake pornography or blackmail combined with so-called doxing. Using people's images to create false, pornographic materials without their consent is one of the most common applications of deepfakes in the context of this type of abuse. On the other hand, doxing involves collecting and using data collected online about other people to embarrass, intimidate, defame, or make the victim feel threatened. This means fake content can be used for blackmailing or publicly shaming victims. Another domain can be indicated in the context of legal and financial threats. Deepfakes can be used to create evidence that can influence court decisions. In addition, generating fake voice or video recordings to impersonate sensitive individuals, such as CEOs (Chief Executive Officers), as part of BEC (Business Email Compromise) attacks can result in financial fraud [7,15,16].

1 DEPARTMENT OF COMPUTER SCIENCE, CZESTOCHOWA UNIVERSITY OF TECHNOLOGY, CZESTOCHOWA, DABROWSKIEGO 69, POLAND, MARIUSZ.KUBANEK@ICIS.PCZ.PL

2 DEPARTMENT OF COMPUTER SCIENCE, CZESTOCHOWA UNIVERSITY OF TECHNOLOGY, CZESTOCHOWA, DABROWSKIEGO 69, POLAND, SABINA.SZYMONIAK@ICIS.PCZ.PL

The last important domain that deepfake technology may affect is public trust. Deepfake technology carries the possibility of creating realistic forgeries. This may undermine the credibility of the media, and as a result, their recipients will begin to question the authenticity of real materials, ultimately leading to a crisis of trust. Considering the ethical aspects of deepfake technology, it should be noted that its development and implementation should consider the protection of individual privacy. Technological safeguards should be introduced to prevent abuse, such as marking deepfake content (watermarking) [9]. This entails the responsibility of creators and users operating in this area. Creators of deepfake tools should be responsible for their potential use, for example, by limiting the accessibility of the technology to people without appropriate permissions. In turn, users should be aware of the moral and legal consequences resulting from the use of deepfakes. Moreover, it is worth considering the introduction of identification systems that indicate that a given material is generated using AI and educating the public, which will aim to increase awareness of the threats and possibilities of identifying deepfake materials. There is also a need to create international legal standards for the use of deepfake technology. Current regulations, such as GDPR [5] or NIS2 [10], do not cover the specific image and text manipulation challenges [2].

Fig.1. Summary of deepfake technology domains where deepfake technology can be used [1]

Mechanisms to protect against deepfake abuse will play an essential role in the coming years. One example is the need to develop deepfake detection technology. We need AI-based tools to analyse multimedia materials anomalies, such as unrealistic eye movements, audio synchronisation errors, or unnatural lighting. Due to the global nature

of the Internet, combating deepfake abuse requires coordination between governments, technology organisations, and international institutions. These, in turn, should be supported by strengthening legal systemsand introducing penalties for using deepfakes for criminal purposes, as well as mechanisms to protect victims [17,3].

Deepfake technology will continue to develop, making detection more difficult and increasing potential threats. In the metaverse era, where digital content will become a key element of interaction, image and text manipulation can greatly impact society. The key challenge will be to find a balance between innovation and responsibility [8].

Deepfake has huge potential, but its irresponsible use poses serious ethical, legal, and social threats. Protection against abuse requires a multidimensional approach, including legal regulations, the development of detection tools, and education of society [4]. Only a responsible approach to this technology will avoid a crisis of trust in the digital era [6].

This article aims to analyse the threats of using deepfake technology and define an ethical framework for its responsible use. We will discuss the current threats resulting from deepfake technology and indicate possible directions for its development in the context of increasing digitisation. We will also propose recommendations for international cooperation and practical technological solutions that can increase security using deepfakes.

References

1. What is deepfake: Ai endangering your cybersecurity? https://www.fortinet. com/resources/cyberglossary/deepfake (2025), accessed: 2025-01-07
2. Akhtar, Z., Pendyala, T.L., Athmakuri, V.S.: Video and audio deepfake datasetsand open issues in deepfake technology: being ahead of the curve. Forensic Sciences 4(3), 289–377 (2024)
3. Cai, Z., Ghosh, S., Adatia, A.P., Hayat, M., Dhall, A., Gedeon, T., Stefanov, K.: Avdeepfake1m: A large-scale llm-driven audio-visual deepfake dataset. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7414–7423 (2024)
4. Dostál, J., Wang, X., Steingartner, W., Nuangchalerm, P.: Digital intelligence - newconcept in context of future of school education. In: Chova, L.G., Martinez, A.L., Torres, I.C. (eds.) 10th International Conference of Education, Research and Innovation (ICERI2017). pp. 3706–3712. ICERI Proceedings, Seville, Spain (November 16–18 2017)
5. 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)
6. Heidari, A., Jafari Navimipour, N., Dag, H., Unal, M.: Deepfake detection usingdeep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 14(2), e1520 (2024)
7. 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)
8. Maheshwari, R.U., Paulchamy, B., Pandey, B.K., Pandey, D.: Enhancing sensingand imaging capabilities through surface plasmon resonance for deepfake image detection. Plasmonics pp. 1–20 (2024)
9. Nadimpalli, A.V., Rattani, A.: Proactive deepfake detection using gan-based visiblewatermarking. ACM Transactions on Multimedia Computing, Communications and Applications 20(11), 1–27 (2024)

10. Schmitz-Berndt, S.: Defining the reporting threshold for a cybersecurity incidentunder the nis directive and the nis 2 directive. Journal of Cybersecurity 9(1), tyad009 (2023)
11. Shi, L., Zhang, J., Ji, Z., Bai, J., Shan, S.: Real face foundation representationlearning for generalized deepfake detection. Pattern Recognition p. 111299 (2024)
12. Steingartner, W., Galinec, D.: The role of categorical structures in infinitesimalcalculus. Journal of Applied Mathematics and Computational Mechanics 12(1), 107–119 (2013). https://doi.org/10.17512/jamcm.2013.1.11
13. Steingartner, W., Novitzká, V.: A new approach to semantics of procedures incategorical terms. In: Novitzká, V., Korečko, S., Szakál, A. (eds.) 2015 IEEE 13th International Scientific Conference on Informatics. pp. 252–257. IEEE, Poprad, Slovakia (November 18–20 2015)
14. Szymoniak, S., Depta, F., Karbowiak, Ł., Kubanek, M.: Trustworthy artificial intelligence methods for users' physical and environmental security: A comprehensive review. Applied Sciences 13(21), 12068 (2023)
15. Szymoniak, S., Siedlecka-Lamch, O.: Securing meetings in d2d iot systems.ETHICOMP 2022 31 (2022)
16. Wang, T., Liao, X., Chow, K.P., Lin, X., Wang, Y.: Deepfake detection: A comprehensive survey from the reliability perspective. ACM Computing Surveys 57(3), 1–35 (2024)
17. Wang, Y., Huang, H.: Audio–visual deepfake detection using articulatory representation learning. Computer Vision and Image Understanding 248, 104133 (2024)

MACHINE LEARNING OR GENERATIVE AI? NEW PERSPECTIVES IN CYBERSECURITY

Mário Marques da Silva[1] [0000-0002-5498-3220], Fernando Correia[2] [0000-0002-9530-1986], Héctor Dave Orrillo Ascama[3] [0000-0003-1555-6821], Laercio Cruvinel Júnior[4] [0000-0002-7672-3243]

Keywords: Cybersecurity and Machine Learning; Generative AI; Hybrid Defence Strategy; Anomaly Detection; Synthetic Data.

Extended Abstract

Introduction

The proliferation of digital systems in the 4th Industrial Revolution has introduced unprecedented cybersecurity challenges. This paper explores the implementation of Generative AI (GEN AI) in the context of cybersecurity, highlighting their applications in data analysis procedures and automatic control systems. By integrating Machine Learning (ML) for real-time detection and GEN AI for simulating advanced attack scenarios, we can detect cyberattacks at an early stage and minimize the impact on the systems functioning.

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 (Samuel, 1959). This paper examines the use of supervised learning methodologies, which rely on labelled datasets (Hinton et al., 2006), and GEN AI, which leverages unstructured data to simulate and preempt cyber threats (OpenAI, 2024). We conclude by discussing the implications for the future of cybersecurity and the anticipated dominance of AI-driven solutions by 2030.

1 Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa, 1169-023 Lisboa, Portugal; Autónoma TechLab, 1169-023 Lisboa, Portugal; Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal; mmsilva@autonoma.pt

2 Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa, 1169-023 Lisboa, Portugal; Autónoma TechLab, 1169-023 Lisboa, Portugal; Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal; fjcorreia@autonoma.pt

3 Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa, 1169-023 Lisboa, Portugal; Autónoma TechLab, 1169-023 Lisboa, Portugal; Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal; horrillo@autonoma.pt

4 Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa, 1169-023 Lisboa, Portugal; Autónoma TechLab, 1169-023 Lisboa, Portugal; Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal; lcruvinel@autonoma.pt

According to Capodieci et al. (Capodieci et al., 2024), there are different types of GEN AI, such as Machine Learning (ML) and Deep Learning (DL), which are a sub-field of AI. While ML depends on the human factor to classify and contextualize data, DL is a system with greater autonomy in processing unstructured data. Moreover, Large Language Models (LLM) represent another type of AI technology, capable of understanding and generating written language (Lee, 2023). 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 (Abadi et al., 2016), and to do that, human interaction is required.

DL, an advanced subset of ML, enhances these capabilities through multi-layer neural networks (Goodfellow et al., 2016). DL is part of unsupervised learning, which is effective in anomaly detection and clustering, identifying previously unseen threats without the need for labelled 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 (OpenAI, 2024). However, its potential misuse, such as generating deepfakes or creating malware variants, highlights the dual-edged nature of this technology (Brynjolfsson, 2014). According to the 2024 report by Capgemini (Capgemini, 2024), in the universe of 1000 organizations analysed, around 45% have encountered deep-fake 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 possible 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 SL 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 (Siam et al., 2025).

Methodology

To explore the integration of ML and GEN AI in cybersecurity, we analyse:

- 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 behaviours, and potential intrusion scenarios, enhancing the ability to predict and mitigate cyber threats.
- A hybrid defence strategy combining the strengths of both paradigms.
- SL can be used as a classier, 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. SL 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 super-vision 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.

Conclusions

The integration of ML and GEN AI offers a robust defense strategy against modern cyber threats. Institutions must invest in both technologies to stay ahead in the cybersecurity arms race. By 2030, it is anticipated that AI will play a critical role in resolving up to 95% of cybersecurity incidents. As the digital landscape continues to evolve, the necessity for adaptive, AI-driven solutions becomes increasingly apparent.

Organizations are increasing their use of AI in various cybersecurity areas. Given the growth in areas related to data analysis, application development or cloud operations, the use of a hybrid defense strategy is essential for the continuity of operations.

Finally, the paradigm shifts in the use of AI, through the increased use of GEN AI combined with ML, is a reality. However, the human component must always be present in the equation to supervise systems' operation and guarantee the presence of ethics in the loop.

References

1. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning, 2016. URL https://arxiv.org/abs/1605.08695.
2. E. Brynjolfsson. The second machine age : work, progress, and prosperity in a time of brilliant technologies. W.W. Norton & Company, 2014.
3. Capgemini. New defenses, new threats: What ai and gen ai bring to cybersecurity. Technical report, Capgemini Research Institute, 2024.
4. N. Capodieci, C. Sanchez-Adames, J. Harris, and U. Tatar. The impact of generative ai and llms on the cybersecurity profession. In 2024 Systems and Information Engineering Design Symposium (SIEDS), pages 448453, 2024. https://doi.org/10.1109/SIEDS61124.2024.10534674.
5. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. The MIT Press, 2016. ISBN 9780262035613.
6. G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. Neural Computation, 18(7):15271554, 2006. https://doi.org/10.1162/neco.2006.18.7.1527.
7. Lee. What are large language models and why are they important? [online] Available: https://blogs.nvidia.com/blog/2023/01/26/whatare-large-language-models-used-for/, Jan. 2023. URL https://blogs.nvidia.com/blog/2023/01/26/what-are-large-language-modelsused-for/.
8. OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

L. Samuel. Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3):210229, 1959. https://doi.org/10.1147/rd.33.0210.

9. A. Siam, M. M. Hassan, and T. Bhuiyan. Artificial intelligence for cybersecurity: A state of the art. In 2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC), pages 1-7, 2025. https://doi.org/10.1109/ICAIC63015.2025.10848980.

CYBERSECURITY AND FOOD SECURITY: A REVIEW OF CYBERATTACK RISKS IN MODERN AGRICULTURE

OLGA SIEDLECKA-LAMCH[1] [0000-0001-9820-6629]

Extended Abstract

Introduction

Food security underpins global sustainability and social welfare, comprising both the availability of food and its nutritional quality and stability across time. The Food and Agriculture Organisation of the United Nations (FAO) characterises food security as a condition 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 [1]

Agriculture 4.0 and 5.0 integrate advanced digital technologies with societies' existing knowledge of crops. Thanks to this, agriculture evolves to the next level of crop methods and agricultural production management. Agriculture 4.0 uses large data sets obtained from numerous devices, sensors, satellite data, and Internet of Things infrastructure in its solutions. This data is often processed using artificial intelligence, giving farmers great opportunities for automation and precise selection of seeds, fertilizers, plant protection products, and irrigation. Agriculture 5.0 goes further by developing IT systems, emphasizing device autonomy, sustainable development, and environmental challenges. Both approaches are intended to increase efficiency, precision, and ensure food security for societies.

In contemporary, digitised agriculture, food security must also consider the stability of the digital systems that support modern food production and delivery. Cyberattacks on agricultural infrastructure threaten both the operational facets of food production, including planting, irrigation, and harvesting, as well as the resilience of global food supply networks. Disruptions to data integrity, communication networks, or essential technology can result in cascading impacts, impacting both the immediate availability and long-term stability of food resources.

In September 2021, NEW Cooperative Inc., an agricultural cooperative in Iowa, was attacked by the Russian-linked group BlackMatter. The attack locked down its computers used to manage food supply chains and animal feeding schedules. BlackMatter threatened to release a terabyte of data it had stolen from thecooperative if it did not receive a $5.9 million ransom. The attack forced the organization to shut down its systems. The incident threatened to disrupt supplies of grain, pork, and poultry [14].

1 CZESTOCHOWA UNIVERSITY OF TECHNOLOGY, CZESTOCHOWA, POLAND, OLGA.SIEDLECKA@ICIS.PCZ.PL

This study analyses the intersection of food security and cybersecurity, highlighting the novel problems it presents and the necessity for innovative solutions to safeguard agricultural systems from increasing technology threats. This work emphasises the significance of safeguarding agricultural systems within the overarching food security framework, underscoring their role as a vital element of global stability through the integration of cybersecurity.

Types of Possible Cyberattacks

In agriculture 4.0 and 5.0, all IoT devices (Edge Devices) are widely used. Attacks on such systems are known in other areas of the economy. Seeing the growing popularity of information technology in agriculture, this area should also be carefully secured. Let's consider possible scenarios.

Intruders can take advantage of the weak security of IoT devices because many of them are poorly secured in agriculture. Users leave default passwords, no encryption, and simple communication protocols are used [7,13]. For example, a hacker can access soil sensors, temperature monitoring, and irrigation control systems and manipulate data, causing improper actions (e.g., excessive irrigation or fertilization).

A botnet attack can turn IoT devices into tools for carrying out DDoS (Distributed Denial of Service) attacks on other systems [2,11].

Using a Man-in-the-Middle (MitM) attack [4], hackers can intercept and modify data transmitted between IoT devices and central management systems, thus disrupting communication in IoT networks [3]. Such attacks can introduce false data during the exchange of information, disrupting farm management systems' operation and leading to incorrect decisions (e.g., incorrect dosage of fertilizers or pesticides).

IoT systems often use wireless networks such as Zigbee, LoRa or Wi-Fi. Radio signal disruption (jamming) [6] can cut off communication between devices, disabling automatic processes on the farm.

Agricultural machinery management systems are often not regularly updated, which leads to attacks using vulnerabilities. One can consider the possibility of injecting malicious code into agricultural machinery control systems (e.g. autonomous tractors), which can lead to their improper operation, crop destruction or even damage to machines [5].

Since data is becoming the most valuable currency in every area of modern life, the possibility of an attack on data should always be considered. Agricultural data is often stored in the cloud and analyzed using platforms such as Climate FieldView or John Deere Operations Center. Attacks on these and similar platforms can result in the theft of agricultural data, such as yield forecasts, soil conditions or field maps, which have high commercial value. Cybercriminals can block access to cloud data through a ransomware attack, encrypting the data and demanding a ransom [10,12]. The farmer then loses the ability to use key information. Hackers can also change soil, weather, or resource consumption data, misleading farmers and causing losses. Agricultural data, such as crop yield, market forecasts or pesticide use data, can be sold to competitors or used for blackmail.

Autonomous agricultural machines (tractors, combines, robots) can also be targeted, for example, by taking control over them, causing them to operate incorrectly or potentially threatening human safety [8].

Central and Eastern Europe is constantly experiencing disruptions to the GPS signal used by agricultural machines for precise navigation. GPS signal disruption (spoofing) can lead to machines' incorrect operation [8].

Social engineering, such as phishing, can also be used to attack modern agricultural systems. Farmers or administrators can be targeted by phishing emails to obtain login details for platforms managing farms, machinery, or IoT systems. Hackers can also impersonate technology/software/machine providers to introduce malware [8].

Defence Techniques

Individual types of attacks can cause direct or further effects on agriculture and supply chains and, consequently, on the food security of societies. Modern agriculture should be appropriately prepared for each type of attack. Unfortunately, it is impossible to predict one universal approach, and in principle, a whole range of security techniques and policies should be created.

Let's consider the weak security of IoT devices; in addition to the requirement of two-factor authentication, it is worth isolating the sensor network from the external network and using end-to-end encryption for communication between devices and the cloud. This isolation can be implemented through network segmentation using VLANs, firewalls, or dedicated gateways that restrict communication to authenticated and authorized endpoints. Some sensors can be integrated using preconfigured frameworks [9], provided that a thorough analysis is conducted to ensure data ownership remains with the farmer and is not shared with third parties.

In the event of jamming threats, distributed communication systems may be employed, and devices can be temporarily put into a low-power sleep mode, periodically checking whether communication has been restored [9].

To counter man-in-the-middle attacks, dynamic sensor configurations can be introduced, dividing devices into clusters that periodically update their membership and communication or authentication methods. Generally, dynamic changes offer a robust defense mechanism for many attack types, particularly considering the limitations inherent in simple sensors.

In the case of GPS signal disruptions, systems incorporating field recognition algorithms and physical markers should be implemented.
Complex and advanced devices should be secured by the standards used in IT.

Conclusions

Agriculture 4.0 and the emerging Agriculture 5.0 are bringing huge benefits in terms of productivity and ecological solutions. However, rapid digital transformation has not been accompanied by an increased awareness of the importance of cybersecurity among those who rely on these technologies. Farmers and agribusinesses often lack the knowledge to recognize, counter, and mitigate cyberattacks.

This study highlights the increasing risks tied to the use of IoT, AI, and cloud platforms in agriculture, particularly due to weak security and the absence of legal protections for farmers. Large corporations often take control of agricultural data, while ownership remains unclear. To address this, tailored regulations inspired by GDPR could help ensure farmers retain data rights and consent to its use. Unregulated data flows, combined with IoT vulnerabilities, pose a serious threat to global food security.

References

1. Berry, Elliot M., et al.: Food security and sustainability: can one exist without theother?. Public health nutrition 18.13, pp. 2293–2302 (2015)
2. Bertino, E., & Islam, N.: Botnets and internet of things security. Computer, 50(2),pp. 76–79 (2017)
3. Cekerevac, Z., Dvorak, Z., Prigoda, L., & Cekerevac, P.: Internet of things and theman-in-the-middle attacks–security and economic risks. MEST Journal, 5(2) (2017)
4. Conti, M., Dragoni, N., & Lesyk, V.: A survey of man in the middle attacks. IEEEcommunications surveys & tutorials, 18(3), pp. 2027–2051 (2016)
5. Goel, S.: A systematic literature review on past attack analysis on industrial controlsystems. Transactions on Emerging Telecommunications Tech., 35(6), e5004 (2024)
6. Ingham, M., Marchang, J., & Bhowmik, D.: IoT security vulnerabilities and predictive signal jamming attack analysis in LoRaWAN. IET information security, 14(4), pp. 368–379 (2020)
7. Meneghello, F., et al.: IoT: Internet of Threats? A Survey of Practical SecurityVulnerabilities in Real IoT Devices. IEEE IoT Journal, vol. 6, no. 5 (2019)
8. Parkinson, S., Ward, P., Wilson, K., & Miller, J.: Cyber threats facing autonomousand connected vehicles: Future challenges. IEEE transactions on intelligent transportation systems, 18(11), pp. 2898–2915 (2017)
9. Pacheco, J., & Hariri, S.: IoT security framework for smart cyber infrastructures.In 2016 IEEE 1st International workshops on Foundations and Applications of selfsystems IEEE (2016)
10. Singh, J., Pasquier, T., Bacon, J., Ko, H., & Eyers, D.: Twenty security considerations for cloud-supported Internet of Things. IEEE Internet of things Journal, 3(3), pp. 269–284 (2015)
11. Szymoniak, S., Piątkowski, J., & Kurkowski, M.: Defense and Security Mechanismsin the Internet of Things: A Review. Applied Sciences, 15(2), 499 (2025)
12. Tejaswi, B., Mannan, M., & Youssef, A.: Security Weaknesses in IoT ManagementPlatforms. IEEE Internet of Things Journal, 11(1), pp. 1572–1588 (2023)
13. Yang Y., Wu L., Yin G., Li L., Zhao H.: A survey on security and privacy issuesin internet-of-things. IEEE Internet of Things Journal, 4(5):1250–8 (2017)
14. New Cooperative Ransomware Attack Timeline: Status Updates, https://www.msspalert.com/news/new-cooperative-ransomware-attack-timelinestatus-updates, last accessed 2025/01/29

ETHICAL CHALLENGES IN THE APPLICATION OF ARTIFICIAL INTELLIGENCE: ANALYSING THE RESPONSIBILITY OF ALGORITHM DECISIONS IN THE CONTEXT OF CYBERSECURITY

Sabina Szymoniak[1] [0000-0003-1148-5691]
Shalini Kesar[2]

Keywords: Artificial Intelligence, Cybersecurity, AI Ethics, Algorithm Transparency, Data Bias.

Extended Abstract

Artificial intelligence (AI) development opens up new possibilities in many areas, such as medicine, transport, education and cybersecurity. At the same time, the dynamic development of this technology raises several ethical challenges, especially in the context of responsibility for decisions made by algorithms. Using AI in critical systems, such as energy infrastructure, finance or personal data protection, requires considering the potential consequences of incorrect algorithm decisions [1,9].

There is a growing use of AI to keep IT systems safe from cyber threats. Some advanced features include machine learning (ML) algorithms and neural networks used to identify unusual network activities and analyze user behaviour and potential threats. Some examples include intrusion detection systems (IDS) and intrusion prevention systems (IPS) that use artificial intelligence to monitor and inspect large volumes of data in real time and identify any potential threat from network traffic [3].

However, certain risks can be associated with the use of AI in cybersecurity. For example, if a wrong threat is identified, it can lead to the stoppage of legal operations, which, on the other hand, can have disastrous security consequences. This is a critical issue in ensuring transparency and auditability of the AI algorithms used in cybersecurity solutions [4].

The major ethical issue in AI is the so-called 'black box' issue. The most complex machine learning algorithms, such as deep neural networks, are often opaque even to their creators. This lack of transparency makes it difficult to determine the cause of actions, thus creating a lack of user trust and hampering the auditing process [6].

In cybersecurity, the lack of transparency can have serious consequences. For example, if an intrusion detection system flags legitimate traffic as a threat, administrators may have difficulty determining why the decision was made. Attackers, in turn, can exploit the lack of auditability to manipulate AI systems to avoid detection. Developing

1 Department of Computer Science, Częstochowa University of Technology, 42-201 Częstochowa, Poland, sabina.szymoniak@icis.pcz.pl
2 Department of Computer Science and Cybersecurity, Southern Utah University, Cedar City, Utah, US, 8472, USA, kesar@suu.edu

interpretable AI methods, which allow for a better understanding of how algorithms work, may solve this problem. AI algorithms are only as good as the data they are trained on. The problem of bias in input data can lead to unfair or discriminatory decisions.

In cybersecurity, biases can lead to unequal treatment of different user groups, for example, through a higher number of false positives for specific geographic locations or device types [7].

The consequences of biases can include:

- Reducing the effectiveness of cybersecurity systems.
- Undermining user trust in data protection systems.
- Risk of legal liability in the event of discrimination resulting from algorithmic decisions.

Data auditing methods and algorithms trained on diversified and representative data sets are necessary to minimise the risk of bias. In addition, systems that monitor the results of decisions made in real-time must be developed [8].

The issue of legal liability for algorithmic decisions is one of the most complex ethical challenges. In the case of AI system errors, it is difficult to determine who should be held responsible—the algorithm's creators, the data providers, or the system's users. In cybersecurity, incorrect AI decisions can have serious consequences, such as personal data breaches, financial losses, or the destabilisation of critical infrastructure [10].

A legal framework that considers the specificity of AI and its applications must be developed. An example is a regulation on liability for damage caused by autonomous AI systems, such as the EU Digital Responsibility Act. A key challenge is balancing innovation and protecting societal interests [5].

To ensure responsible use of AI, it is necessary to implement a multi-dimensional approach, including:

1. Transparency: development of Explainable AI technology.
2. Education: raising users' awareness about AI's possibilities and limitations.
3. Legal regulations: introducing clear rules of responsibility for algorithm decisions.
4. Auditing and monitoring: regular checking of the effectiveness of AI systems and their potential biases.

In cybersecurity, it is also essential to develop tools for detecting algorithm abuse and manipulation and supporting cooperation between governments, technology companies, and international organisations [2,5].

AI has enormous potential for developing cybersecurity, but its irresponsible use can lead to serious ethical and legal threats. Transparency, responsibility, and education are key elements that will minimise the risks associated with AI. Only a comprehensive approach, combining technological, ethical, and legal aspects, will allow for the safe and responsible use of artificial intelligence in critical systems such as cybersecurity [11].

The full paper will investigate the multifold ethical, legal, and technological problems of applying artificial intelligence in cybersecurity. It will discuss the possible dangers of using AI-based solutions, which include bias in data, lack of transparency in

decision-making (the black box problem) and consequences of incorrect threat identification in vital systems. The paper will also review the impacts of these challenges, which can be damaging and include decreased system effectiveness, lost user trust, and possible legal consequences. The paper will suggest a multi-dimensional approach to these issues, including transparency through explainable AI, legal frameworks for accountability, AI literacy and education on its limitations, and real-time auditing. Hence, this paper seeks to plot a course between technological innovation and ethical responsibility to deploy safe, trustworthy AI in cybersecurity systems.

A conceptual perspective informs this paper, and an analysis of the surrounding literature is based upon a review of scientific research and current legal regulations, including the EU Digital Responsibility Act. It also examines two key case studies involving AI in cybersecurity to identify significant challenges and possible solutions. This study analyses the ethical challenges associated with using AI in cybersecurity. A particular emphasis is placed on challenges associated with the lack of transparency to algorithms (the so-called "black box" problem), and bias in training data, its impact on AI decision-making, and legal liability regarding AI decisions related to data breaches and cyber threats.

Additionally, this study will suggest a technological, legal, and ethical framework that will promote better practices in utilising AI in this capacity. Unlike previous studies that primarily approach the issue through a technical lens, this analysis is interdisciplinary and collaboratively incorporates ethical, legal, and technological considerations. The paper also provides concrete suggestions regarding the use of Explanation AI (XAI) and mechanisms for real-time auditing of AI decision-making to assign and enhance transparency and accountability of algorithms. Responsible AI use in cybersecurity is central to creating appropriate legal regulations. An example of this action is the EU Digital Responsibility Act, which assigns responsibilities to creators of AI and users of AI surrounding transparency and accountability regarding algorithm-based decisions. The paper will analyse its key provisions and potential implications for cybersecurity.

References

1. Awadallah, A., Eledlebi, K., Zemerly, J., Puthal, D., Damiani, E., Taha, K., ... & Yeun, C. Y. (2024). Artificial intelligence-based cybersecurity for the metaverse: research challenges and opportunities. IEEE Communications Surveys & Tutorials, (2024).
2. Galinec, D., Steingartner, W., Combining cybersecurity and cyber defense to achieve cyber resilience. In: 2017 IEEE 14th International Scientific Conference on Informatics. IEEE, p. 87-93, (2017).
3. González, A. L., Moreno, M., Román, A. C. M., Fernández, Y. H., & Pérez, N. C. Ethics in Artificial Intelligence: an Approach to Cybersecurity. Inteligencia Artificial, 27(73), 38-54, (2024).
4. Hashmi, E., Yamin, M. M., & Yayilgan, S. Y., Securing tomorrow: a comprehensive survey on the synergy of Artificial Intelligence and information security. AI and Ethics, 1-19, (2024).
5. Kun, Eyup. Searching for the appropriate legal basis for personal data processing for cybersecurity purposes under the NIS 2 Directive: Legal obligation and/or legitimate interest?. Computer Law & Security Review, 56: 106098, (2025).
6. Malatji, M., & Tolah, A., Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. AI and Ethics, 1-28, (2024).
7. Perumal, A. P., Chintale, P., Molleti, R., & Desaboyina, G. Risk Assessment of Artificial Intelligence Systems in Cybersecurity. American Journal of Science and Learning for Development, 3(7), 49-60, (2024).

8. Upadhayay, Y., & Sharma, R., Cybersecurity And Legal Considerations In AI Applications For National Security. Educational Administration: Theory and Practice, 30(5), 11797-11803, (2024).
9. Szymoniak, S., Kubanek, M., Biometry-based verification system with symmetric key generation method for internet of things environments. Scientific Reports, 15.1: 5464, (2025).
10. 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, 2021. p. 513-525.
11. 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).

CYBERSECURITY CURRICULUM TO INCLUDE ETHICS AND PRIVACY

SHALINI KESAR[1]

Keywords: Ethics, Privacy, NSEE framework, Cybersecurity Curriculum.

Extended Abstract

This paper is part of on-going research on how to include ethics and privacy in curriculum designed for cybersecurity students [2]. It discusses and addresses how the author modified the existing established framework designed for proofing students with experiential learning. The author uses a framework specially designed for educational experiential learning for online students, the right ethical practices of National Society for Experiential Education [3] This paper also highlights the motivation and the previous work of the author with the same framework in a different discipline [1]. It sheds light on how the author designed the classes and how it has been beneficial in creating an interactive online presence for students. The approach used in this paper can be beneficial to other educators who can use this approach in the classroom to build on the examples that can provide learning with real-world experiences and active participation among students. It also promotes experiential learning, which involves engaging students in hands-on activities and projects, encouraging them to apply knowledge and skills in real-world contexts. Educators can create curriculum that emphasizes experiential learning as a core element of education, advocating for hands-on experiences and reflection to foster deeper understanding and engagement in students.

Founded in 1971, the Society for Experiential Education (SEE) is the premier, nonprofit membership organization composed of a global community of researchers, practitioners, and thought leaders who are committed to the establishment of effective methods of experiential education as fundamental to the development of knowledge, skills and attitudes that empower learners and promote the common good [3]. The framework consists of eight principles linked with good practices. In this paper, the instructor (author) as part of an experiential learning activity discusses the modification implemented when designing the curriculum. The goal was that this experience and the learning will add value to the fundamentals of creating online cybersecurity training as part of a group project. Consequently, ensuring both the quality of the learning experience and of the work produced by the students, and in building an assignment that underlies the pedagogy of experiential education. Although the NSEE framework was used as the main thought process was very different when designing the project. It considered the framework as well as the research regarding team projects and the im

1 DEPARTMENT OF COMPUTER SCIENCE AND CYBERSECURITY, SOUTHERN UTAH UNIVERSITY, CEDAR CITY, UTAH, US, 8472, USA, KESAR@SUU.EDU

portance of training in cybersecurity. This is because this style of pedagogy will provide an experiential learning education environment, which will better prepare the student to face challenges in the ever-evolving cybersecurity field.

This paper is part of ongoing research and not meant to be a short-term project, rather the intention to continue the research as a long-term collaborative that will continue to refine and revise its suggestions in cybersecurity pedagogy to prepare students to meet the high demand of the workforce in cybersecurity. When designing the assignments and the course, the author met with industry members to outline a few key points to consider in the design. For example:

- What will the students in teams do?
- How will industry members contribute as guests?
- Will ethics and privacy aspects be incorporated in the assignment or a separate assignment?
- What are specific issues in experiential education assignments should be addressed?

While developing the curriculum, various studies were considered, including author's previous published research. The best standards and Guiding Principles of Ethical Practice by the National Society for Experiential Education (NSEE) was used as a framework with the NSEE Guiding Principles of Ethical Practices to develop the pedagogy to teach ethics and professional as part of an experiential education. This paper describes how the instructor included ethics and professionalism in this team project. The eight principles are exampled below.

Intention: In this principle, all parties must outline a clear vision on the reason which this experience was chosen and why experience is the chosen approach to learning. This principle was used to develop assignments that included induvial reports and teamwork with an ethical and private aspect of cybersecurity, that is often overlooked. Furthermore, it focuses on the purposefulness that enables experience to become knowledge and, as such, is deeper than the goals, objectives, and activities that define the experience. When designing the course, the author's intention was to introduce students to organizations/Associations who are committed to educating both private users and large businesses on cybersecurity safety measures. For example, SANs Institute, ISACA, Information Systems Security Association.

Preparedness and Planning: The main objective of this principle was to ensure the students, both traditional and non-traditional from different educational backgrounds have a experiential learning that encapsulates the new emerging trends in the cybersecurity field. The topics in this course were planned and research to ensure they aligned with the identified intentions, adhering to them as goals and objectives of the overall degree. The hands-on activities included research and critical thinking as a key component while problem solving. Some of the reading material from associations such as Information Systems Association (CISSA), who focus on ethics and privacy were used as resources. This is because The ISSA Code of Ethics is a prerequisite for membership, reinforcing their commitment to ethical principles. The ISSA is a global organization dedicated to advancing information security and has a strong focus on ethical conduct. They also offer student memberships at a reduced rate, allowing students to access the same

resources and benefits as full members, including access to the members' area, event discounts, and the ISSA Journal.

Authenticity: It is important for the students to have an experience that is in a real-world context. For this class, an assignment on cyber law and ethics was developed, as a scenario-based case. Such topics help students to be better equipped with the necessary ethical framework to navigate complex situations in the digital world.

Reflection: NSEE refers to Reflection as an element that transforms simple experience to a learning experience. With this principle in mind, the assignment was designed so that knowledge can be discovered and internalized as the student's research, test assumptions and hypotheses about the outcomes of decisions and actions taken in context of cybersecurity training. It also gave them an opportunity to reflect and understand the societal implications of their actions and make responsible decisions when handling sensitive data, balancing security needs with individual privacy rights, and upholding ethical standards within their professional roles. This process in the assignment comprises of a report writing and presentations at conference and as a final exam. This, according to NSEE, is integral to all phases of experiential learning, from identifying intention and choosing the experience, to considering preconceptions and observing how they change as the experience unfolds. As part of reflection, students were encouraged to present their findings at conferences within the university and beyond. Some of the students were invited to be part of the panel session conducted by the author during the Cyber Awareness Month and attend the national conference Women in Cybersecurity.

Orientation and Training: The students were required to discuss their learnings as a part of an interactive assignment on Canvas. It allowed students to experience and learn about each other and about the context and environment in every changing field. Monitoring and Continuous Improvement: In this principle, it is important to note that any learning activity designed should be dynamic and changing. The various assignments with different topics provided the richest learning possible to the students. Students also had to write a self-reflection on their own progress as well as their team members. This feedback process relates to learning intentions and quality objectives. Consequently, this allows the structure of the experience to be sufficiently flexible that permitted changes in response to what that feedback suggests. Subsequently, monitoring and continuous improvement represent formative evaluation tools.

Assessment and Evaluation: Assessment is a means to develop and refine the specific learning goals and quality objectives identified during the planning stages of the experience. Whereas evaluation provides comprehensive data about the experiential process as a whole and whether it has met the intentions which suggested it. Based on the NSEE definitions, the outcomes and processes of assignments of the project included systematically reports, presentations, and self- reflection that were linked with the initial intentions.

Acknowledgment: At the end of the project, the assignment also included that students recognize the lessons learned, recognition of learning and impact occur throughout the experience by way of the reflective and monitoring processes and through reporting, documentation and sharing of accomplishments. All the students' and instructor's experience were noted and included in the lessons learned and reflected in the recognition of progress and accomplishment. Given that this was part of on-going research where other projects used NSEE's framework, the

lessons learned from culminating documentation and the impact of these projects were part of designing as well as helped to provide closure and sustainability to the experience.

References

1. Kesar, S., and Pollard, J. (2021) "Cultivating an Empathic Learning Pedagogy: Experiential Project Management", in Normal Technology Ethic Proceedings of the ETHICOMP* 2021, Coords. Mario Arias Oliva, Jorge Pelegrín Borondo, Kiyoshi Murata, Ana María Lara Palma, Universidad de La Rioja, 257-259. https://dialnet.unirioja.es/servlet/libro?codigo=824595
2. Kesar, S., and Pollard, J, (2020) Lesson Learned from Experiential Project Management Learning Pedagogy", in Paradigm Shifts in ICT Ethics Proceedings of the ETHICOMP* 2020, Coords. Mario Arias Oliva, Jorge Pelegrín Borondo, Kiyoshi Murata, Ana María Lara Palma, Universidad de La Rioja, 99-100.
3. The National Society (2009). Guiding Principals of Ethical Practices. https://www.nih.gov/health- information/nih-clinical-research-trials-you/guiding-principles-ethical-research.

CYBERSECURITY, BEST PRACTICES: A COMPREHENSIVE GUIDE

PEDRO RAMOS BRANDAO[1]

Keywords: Cybersecurity, incident, artificial intelligence.

Extended Abstract

In an age where digital infrastructures support nearly every sector of modern society, the need for robust and adaptable cybersecurity frameworks has become indisputable. Pedro Brandao's Cybersecurity Best Practices: A Comprehensive Guide offers a criti- cal, detailed, and application-oriented examination of how organizations, including small and medium-sized enterprises (SMEs), can bolster their defenses against the ex- panding and evolving range of cyber threats. Drawing from both academic theory and professional practice, the guide integrates strategic planning, technological implemen- tation, policy development, and human factor management into a cohesive cybersecu- rity blueprint.

The abstract introduces the work's central premise: cybersecurity must be viewed not as a static compliance checklist, but as a dynamic and continuous process that is deeply integrated into the fabric of every organization. As cybercrime becomes increas- ingly automated, politically motivated, and financially lucrative, threats like ransom- ware, phishing, and insider attacks have evolved in both scale and sophistication. This paper presents a scalable and adaptable response framework designed to meet the needs of organizations with varying resources and maturity levels.

At the foundation lies risk assessment and management, which the author identifies as the cornerstone of effective cybersecurity. Brandão proposes that risk identification should start with a comprehensive audit of digital assets, including intellectual property, financial data, operational systems, and sensitive personal records. Afterwards, risks are prioritized based on likelihood and potential impact, leading to the distribution of mitigation responsibilities among specialized staff. The importance of shared responsi- bility and decentralized ownership of risk treatments is highlighted as a strategy for resilience and sustainability.

The author also advocates for designing and enforcing information security policies that are continually updated to reflect new threats, organizational changes, and evolving regulatory requirements. The guide offers a modular methodology for developing such policies, including drafting, stakeholder engagement, internal audits, and staff training.

1 INSTITUTO SUPERIOR DE TECNOLOGIAS AVANÇADAS, LISBOA, PORTUGAL PEDRO.BRANDAO@ISTEC.PT

A particular strength of the guide is its integration of international standards such as ISO/IEC 27001 and NIST SP 800-53, which serve as both benchmarks and strategic tools for governance.

The chapter on the cyber threat landscape offers a typology of modern threats, rang- ing from advanced persistent threats (APTs) to industrial espionage and zero-day ex- ploits. It details how attackers leverage automation, AI, and social engineering to breach even well-protected systems. Importantly, the guide underscores the risks con- nected to "gray zone" threats—acts that fall short of conventional warfare but are aimed at undermining digital sovereignty and economic stability.

In practical terms, network security involves recommendations for layered defense models, firewall configuration, intrusion detection systems (IDS), and endpoint seg- mentation. The guide provides empirical evidence showing that optimal configuration and integration, rather than the latest technology, can significantly enhance threat de- tection and containment. This understanding is especially important for SMEs operat- ing with limited cybersecurity budgets.

Data protection mechanisms, such as encryption and data masking, are examined with technical clarity. The paper details how data should be secured both at rest and in transit and explains how cryptographic methods should be integrated into broader data lifecycle management systems. It emphasizes data minimization, access control, and auditability, all of which are vital components of GDPR compliance and wider privacy legislation.

The section on incident response outlines a comprehensive protocol adapted from the Incident Command System (ICS) that details roles, actions, and communication flows during a security breach. From initial detection and containment to post-incident analysis, the guide ensures that readers understand both the strategic and tactical di- mensions of an effective response.

The human aspect of cybersecurity is a recurring theme, culminating in the section on cybersecurity awareness training. The author convincingly argues that technical in- frastructure is only as secure as its weakest human operator. Therefore, it is essential to train staff to recognize phishing attempts, social engineering tactics, and data handling protocols. The guide includes models for ongoing training, gamified learning, and sim- ulated attack scenarios.

Legal compliance is examined through the lens of the GDPR and similar frame- works. The paper discusses how regulatory requirements have transformed corporate approaches to data governance, privacy, and breach notification. Key GDPR princi- ples—such as privacy by design, data subject rights, and extraterritorial enforcement—are explored, along with practical guidance on implementation.

Cybersecurity Best Practices

The paper concludes with an exploration of emerging technologies in cybersecurity, specifically artificial intelligence (AI) and machine learning. These technologies are presented not only as defense enhancers—utilized in anomaly detection and threat pre- diction—but also as offensive tools used by adversaries. This dual-use nature under- scores the significance of ethics and accountability in AI deployment.

Furthermore, real-world case studies involving critical infrastructure, such as SCADA systems and healthcare networks, offer tangible examples of how best practices are implemented under pressure. These scenarios emphasize the significance of cross- functional coordination, real-time monitoring, and forensic readiness.

Ultimately, this guide emphasizes that cybersecurity is not merely a technical issue but a multifaceted challenge requiring leadership, policy, training, and continuous improvement. By clarifying complex concepts and offering actionable recommendations, it empowers readers with both the mindset and the tools to cultivate a culture of cyber resilience.

References

1. J. M. Borky and T. H. Bradley, Protecting Information with Cybersecurity, National Academies Press, 2018.
2. M. Altaher Ben Naseir et al., "Contextualising the National Cyber Security Capacity in an Unstable Environment: A Spring Land Case Study," Information and Computer Security, vol. 27, no. 3, pp. 349–366, 2019.
3. M. Schmitt, "Mobile Security for the Modern CEO: Attacks, Mitigations, and Future Trends," Cybersecurity Journal, vol. 14, no. 1, pp. 22–30, 2022.
4. U. Kaila and L. Nyman, Information Security Best Practices: First Steps for Startups and SMEs, Finnish Information Security Council, 2018.
5. M. B. Murtaza, "Developing an IT Risk Assessment Framework," Journal of Information Systems, vol. 21, no. 1, pp. 63–75, 2007.
6. J. R. C. Nurse, S. Creese, and D. De Roure, "Security Risk Assessment in Internet of Things Systems," IEEE Security & Privacy, vol. 15, no. 5, pp. 24–30, 2017.
7. M. Alshaikh et al., "Information Security Policy: A Management Practice Perspective," Computers & Security, vol. 59, pp. 1–14, 2016.
8. M. S. Weiss, "Network Defense: The Attacks of Today and How Can We Improve?," Communications of the ACM, vol. 52, no. 12, pp. 40–47, 2009.
9. H. Cavusoglu et al., "Assessing the Value of Network Security Technologies," Information Systems Research, vol. 27, no. 3, pp. 585–602, 2016.
10. A. Olajide and T. Olarewaju, "Application of Data Masking in Achieving Information Privacy," International Journal of Information Management, vol. 34, no. 4, pp. 563–569, 2014.

THE ETHICAL DILEMMAS OF AI-POWERED CYBERSECURITY: BALANCING PRIVACY AND PROTECTION

Nikita Godinho[1]

Keywords: Artificial Intelligence, Cybersecurity, Ethics, Privacy, Bias, Regulation, Surveillance.

Extended Abstract

Introduction

By 2025, cybercrime is expected to cost the global economy $10.5 trillion annually. As organizations and governments increasingly rely on AI-powered cybersecurity tools to detect, analyze, and prevent cyberattacks, ethical concerns about privacy, bias, and accountability are coming to the forefront. These tools enhance security by identifying suspicious activity, automating responses, and adapting to new threats. However, the integration of AI into cybersecurity raises ethical questions that must be addressed. This essay explores the ethical challenges of AI-powered cybersecurity, focusing on four key areas: mass surveillance and privacy violations, bias and discrimination in AI systems, the lack of transparency and accountability in AI decision-making, and the ethical dilemmas of AI-powered cyberattacks.

A striking example of these challenges happened in 2024, when scammers used deepfake video and voice to impersonate top executives at Arup, a British engineering company. During a fake video call, they convinced an employee in the Hong Kong office to transfer over $25 million to fraudulent accounts. The attackers used AI to mimic the appearance and voice of Arup's CFO and other staff, making the call look and sound real. The scam wasn't just a technical breach , it exploited trust and blurred the line between genuine human communication and AI deception. This case shows how AI tools can be used not only to protect systems but also to undermine them in deeply personal and costly ways.

Ethical Perspectives

- Deontology: Do AI systems uphold moral obligations and protect fundamental rights, such as privacy?
- Utilitarianism: Do these tools maximize overall security benefits while minimizing harm to individuals?

1 Universidade Autónoma

- Virtue Ethics: Are AI cybersecurity initiatives guided by principles of responsibility, transparency, and fairness?

Understanding How AI Security Models Work

AI-powered cybersecurity systems detect and respond to cyber threats in real time using various machine learning approaches. These models can be categorized as follows:

- Supervised Learning: Trained on labeled cyber threat data to recognize known attack types.
- Unsupervised Learning: Detects patterns or anomalies without predefined labels.
- Reinforcement Learning: Learns from real-world outcomes to adaptively improve.

Despite these benefits, AI security models have limitations:

- Bias in data training can lead to unfair targeting.
- Lack of transparency makes it difficult to explain or challenge decisions.
- False positives/negatives can result in misclassifying threats or flagging legitimate behavior.

Mitigation strategies:

- Conduct regular bias audits to identify and address biased behaviors in AI models.
- Develop AI models with interpretable decision-making processes to improve transparency.
- Ensure human oversight in cybersecurity decision-making to maintain accountability through human-AI collaboration.

Case Studies

NSA PRISM-Privacy vs. Protection

The NSA's PRISM program collected vast user data without consent under the banner of national security. From a deontological perspective, it violated privacy rights. From a utilitarian one, it questioned whether collective safety justified such intrusions. The case clearly illustrates the ethical trade-off between national security and personal privacy, especially when oversight is lacking. Over time, unchecked data collection can normalize surveillance and erode public trust.

Facial Recognition Bias

Studies show facial recognition systems misidentify people of color at much higher rates. This stems from unbalanced training data and a lack of transparency. Without proper oversight, these systems can reinforce discrimination in security contexts. Regular audits, diverse datasets, and human intervention are essential.

Solutions for Ethical AI Cybersecurity

Regulatory Strategies:

- Transparency Mandates: AI systems must explain their decisions clearly.
- Data Protection Laws: Extend rules like GDPR to AI security systems.

- Oversight Mechanisms: Regulators must monitor AI use to ensure accountability.

Technical Solutions:

- Bias Detection: Use tools to detect and reduce bias in models.
- Interpretable Outputs: Build explainable systems to avoid black-box risks.
- Human-AI Collaboration: Combine automation with human review to improve ethical oversight.

Industry Best Practices:

- Ethical AI Certification: Require external validation of responsible AI practices.
- Diverse Development Teams: Involve people from varied backgrounds in AI design.
- User Control: Give users more say over how security AI affects them.

Implementation Challenges

Even though these solutions make sense on paper, actually applying them at scale is a different story. Things like bias audits, human oversight, and clear rules take time, money, and the right people , and not every organization has that. Smaller organizations, in particular, may struggle to keep pace with evolving standards or invest in the resources needed to implement these safeguards properly. Even bigger companies can face challenges staying consistent.Without shared standards across countries or sectors,

these efforts can easily fall apart or end up being more symbolic than effective. So while the ideas are strong, turning them into real-world practice is still a big challenge.

Conclusion

AI-powered cybersecurity offers immense potential but also possesses significant ethical challenges. While AI enhances security efficiency, responsible implementation is necessary to prevent privacy violations, bias, and a lack of accountability. Ethical theories such as deontology, utilitarianism, and virtue ethics provide structured frameworks for addressing these challenges. Governments, businesses, and researchers must collaborate to establish transparent regulations, develop advanced bias-reduction techniques, and promote responsible AI development. By embedding justice, accountability, and oversight into AI-driven cybersecurity, we can protect digital assets without compromising fundamental human rights.

References

1. Greenwald, G., MacAskill, E.: NSA Prism program taps into user data of Apple, Google and others. The Guardian (2013). https://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data
2. Buolamwini, J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research, 81, 1–15 (2018). http://proceedings.mlr.press/v81/buolamwini18a.html

3. European Parliament: General Data Protection Regulation (GDPR), Regulation (EU) 2016/679. https://gdpr.eu/, last accessed 2 April 2025
4. World Economic Forum: Cybercrime: Lessons learned from a $25m deepfake attack. https://www.weforum.org/stories/2025/02/deepfake-ai-cybercrime-arup/, last accessed 2 April 2025
5. Fortune: A deepfake 'CFO' tricked British design firm Arup in $25 million fraud. https://fortune.com/europe/2024/05/17/arup-deepfake-fraud-scam-victim-hong-kong-25-million-cfo/, last accessed 2 April 2025
6. The Guardian: UK engineering firm Arup falls victim to £20m deepfake scam. https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video, last accessed 2 April 2025
7. American Civil Liberties Union (ACLU): The NSA Continues to Violate Americans' Internet Privacy Rights. https://www.aclu.org/news/national-security/nsa-continues-violate-americans-internet-privacy, last accessed 2 April 2025
8. European Commission: Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (AI Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206, last accessed 2 April 2025
9. Garvie, C.: The Perpetual Line-up: Unregulated Police Face Recognition in America. Georgetown Law Center on Privacy & Technology (2016). https://www.perpetuallineup.org/
10. (ISC)²: The Ethical Dilemmas of AI in Cybersecurity. https://www.isc2.org/Insights/2024/01/The-Ethical-Dilemmas-of-AI-in-Cybersecurity, last accessed 2 April 2025

CHATGPT INTEGRATION WITH THE E-CITIZENS PORTAL: A PROTOTYPE OF A WEB APPLICATION TO IMPROVE ACCESS TO HEALTH INFORMATION

Natalija Zeko[1], Darko Galinec[2], William Steingartner[3]

Keywords: Artificial intelligence, ChatGPT, healthcare, virtual assistant, web application, implementation, PHP, MySQL

Extended abstract

This paper deals with the application of artificial intelligence, especially ChatGPT, in the medical sector through the web application e-Doctor. The paper of two parts: theoretical and practical. The theoretical part provides an insight into the basic concepts of artificial intelligence, its history, types, and applications, with an emphasis on the medical sector. Modern technology is still some way from achieving "third wave AI" capabilities, also known as "artificial general intelligence" or "strong AI" [1, 2].

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 [2, 3]. This notion of AI transforming the medical profession is also echoed in literature, such as Robin Cook's novel Cell (2014), where a smartphone-based system called iDoc acts as a fully capable personal physician, outperforming human doctors. This fictional scenario reflects real-world discussions about the evolving role of AI in healthcare, as explored through the e-Doctor application in this paper.

The paper includes various diagrams to illustrate the system structure, including business processes, data models, and user activities. The practical part of the paper focuses on the development of a prototype web application, e-Doctor, which utilizes ChatGPT. ChatGPT acts as a virtual assistant for users, answering any medical-related questions they may have. The web application offers functionalities such as viewing medical data, sending questions to the virtual assistant, and generating personalized meal plans. The main goal of the web application is to reduce the administrative workload on medical workers and improve the quality of medical services [4, 5, 6]. With the development of technology and digital systems, the daily lifestyle is undergoing significant changes. As one of the most significant technological advances of today, dating back to the mid-20th century, artificial intelligence today has applications in various indus

1 Zagreb University of Applied Sciences (Croatia), zekonatalija@gmail.com
2 Zagreb University of Applied Sciences (Croatia), darko.galinec@tvz.hr
3 Technical University of Košice (Slovakia), william.steingartner@tuke.sk

tries, especially in healthcare. Digitalization enables faster and more efficient access to information, and the advancement of artificial intelligence brings new opportunities to improve health services [7]. The motivation for creating this paper stems from the incredible possibilities that artificial intelligence brings, enabling complex problems to be solved and new pathways to be discovered in the fight against disease. Previously, we were not aware that this technology would become a key tool in improving diagnostics, treatment and access to health information [8]. In this paper, the goal is to show 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 insight into how artificial intelligence can improve access to health information, enabling faster and more accurate responses to patients [9, 10]. The main goal of the web application is to make it easier for patients to access health information, and to reduce the administrative burden on healthcare professionals [11]. During the preparation of the paper, various websites, books, professional and scientific articles with appropriate topics were used for the purpose of writing the paper. The paper consists of a theoretical and a practical part. At the very beginning of the paper, the definition of artificial intelligence is explained, as well as its history, types, branches and applications. After that, the basic characteristics of ChatGPT and e-Citizens are presented. After applying the knowledge about the characteristics of e-Citizens, a visual representation of the diagram is displayed, which will be crucial when creating the e-Doctor web application. The diagrams that will be shown are the use case diagram, activity diagram, work diagram, business process model, sequence diagram, data model diagrams [12, 13]. During the practical implementation, the technologies that were used in the work are listed. Finally, the functionality of the prototype of the e-Doctor web application for improving access to health information is presented.

When creating application system, several programming languages and technologies were used to ensure the functionality, interactivity, and adaptability of the application. For the basic structure and organization of the content on the website, HTML was used. HyperText Markup Language). CSS Cascading Style Sheets are used to style elements, i.e. to customize the color, font, and appearance of a page. Bootstrap is a front-end framework that offers predefined components and styles, which made it possible to quickly create a user interface. A MySQL database was used to store and manage data. In the background part of the application, PHP is used. Hypertext Preprocessor) Symfony 6 framework that enabled the overall functionality of the application such as database work, 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 asked. This way leads to a better and faster user experience. 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 wargame scenarios. Drawing upon Game Theory and machine learning to balance defenders' advantage, work on modeling and

lation 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.

The main goal was to show the application of artificial intelligence in today's time. 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 paper. In the practical part, the focus is on ChatGPT, an advanced language model developed by OpenAI [14, 15, 16]. In the paper, ChatGPT was used as a virtual assistant in the e-Doctor app as a virtual assistant to provide health information and advice that facilitates communication between patient and doctor. Through diagrams of use cases, activities and business processes, key processes within the application are visually shown. The development of a prototype of the e-Doctor web application is an important step in the modernization of health services, providing patients and healthcare professionals with faster and easier access to information. The prototype thus shows the potential for improving the quality of health care, 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 Application Programming Interface (API) connection to connect the application to the health system database. As far as security upgrades are concerned, it is necessary to further ensure data privacy, which is achieved by replacing the current authentication system, with the system of the National Identification and Authentication System NIAS and e-Authorization. According to the General Data Protection Regulation (GDPR), a regular revision of safety requirements is required. Users with certain disabilities such as blindness and visual impairment will try to improve accessibility by enlarging fonts, using 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 the video call system. The final activity involves conducting user experience testing to identify potential issues or improvements. Analyzing the data [9] 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 [17, 18, 19, 20, 21, 22] 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.

Acknowledgment: 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] 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, vol. 183, no. 6, June 2023, pp. 589–596. doi:10.1001/JAMAINTERNMED.2023.1838.

[2] Betz, Sunny. "7 Types of Artificial Intelligence." Built In. Retrieved May 2, 2024. https://builtin.com/artificial-intelligence/types-of-artificial-intelligence.

[3] Bohnenberger, T. AI & The Industry. epubli. Retrieved May 2, 2024. https://www.google.hr/books/edition/AI_The_Industry/uzboEAAAQBAJ.

[4] Bowman, Jeremy. "How Is AI Used in Healthcare?" The Motley Fool. Accessed May 6, 2024. https://www.fool.com/investing/stock-market/market-sectors/information-technology/ai-stocks/ai-in-healthcare/.

[5] Briganti, G., and Le Moine, O. "Artificial Intelligence in Medicine: Today and Tomorrow." Frontiers in Medicine, vol. 7, Feb. 2020. doi:10.3389/fmed.2020.00027.

[6] Copeland, B. J. "Artificial Intelligence - Machine Learning, Robotics, Algorithms." Britannica. Retrieved May 1, 2024. https://www.britannica.com/science/history-of-artificial-intelligence.

[7] Das, Suvankar. "What are the Components of AI?" ellow.io. Accessed May 2, 2024. https://ellow.io/components-of-ai/.

[8] Digital Adoption Team. "The Most Important Branches of Artificial Intelligence You Need To Know." Accessed May 2, 2024. https://www.digitaladoption.com/branches-of-artificial-intelligence/.

[9] 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.

[10] Gulati, Ashish. "The History of Artificial Intelligence (AI): A Detailed Guide." KnowledgeHut. Retrieved May 2, 2024. https://www.knowledgehut.com/blog/data-science/history-of-ai.

[11]Ingram, Andrew. ChatGPT. Greenbooks editore. Accessed May 8, 2024. https://www.google.hr/books/edition/ChatGPT/cGrfEAAAQBAJ.

[12] Johnson, C. K., R. W. Van Tassel, A. Rogers, T. Shade, and K. Ferguson. "Adversarial Cognitive Engineering (ACE) and Defensive Cybersecurity: Leveraging Attacker Decision-Making Heuristics in a Cybersecurity Task." Proceedings of the 57th Hawaii International Conference on System Sciences, 2024. https://www.researchgate.net/publication/380209604.

[13] 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.

[14] Malik, Eddy. "Artificial Intelligence (AI) and ChatGPT: History and Timelines." Office Timeline. Accessed May 8, 2024. https://www.officetimeline.com/blog/artificial-intelligence-ai-and-chatgpt-history-and-timelines.

[15] Marr, Bernard. "How Is AI Used In Healthcare - 5 Powerful Real-World Examples That Show The Latest Advances." Forbes. Accessed May 7, 2024. 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/.

[16] Mijwil, M. M. "History of Artificial Intelligence." vol. 3, Jan. 2015. doi:10.13140/RG.2.2.16418.15046.

[17] Pandey, Vibha. Artificial Intelligence for Students: A Comprehensive Overview of AI's Foundation, Applicability, and Innovation (English Edition). Bpb Publications. Retrieved May 1, 2024. https://books.google.hr/books?id=ptq1EAAAQBAJ.

[18] Roser, Max. "The Brief History of Artificial Intelligence: The World Has Changed Fast — What Might Be Next?" Our World in Data. Retrieved May 2, 2024. https://ourworldindata.org/brief-history-of-ai.

IN THE HANDS OF AI: A CYBERSECURITY PHILOSOPHICAL APPROACH

Nuno Silva[1] [0000-0003-0157-0710], Pedro Brandao[2] [0000-0001-6351-6272]

Keywords: AI, cybersecurity ecosystems, philosophical, practical landscape.

Extended Abstract

The intersection of artificial intelligence (AI) and cybersecurity presents a complex philosophical and practical landscape, characterized by both opportunities and challenges. This research is an exploration of this topic from a cybersecurity philosophical approach. Cybersecurity systems are interconnected socio-technical ecosystems influenced by external forces like regulatory frameworks and societal norms. Machine learning algorithms improve threat detection by identifying unusual patterns that could indicate breaches which no person can do in a timely manner. Meanwhile, involving diverse stakeholders in designing AI-driven cybersecurity systems can help align technology with societal values and ethical principles. Human sense and sensitivity to cybersecurity is under a revolutionary transformation at the hands of AI. For example, when AI autonomously makes decisions - such as blocking IPs or quarantining files - questions arise about accountability. Determining responsibility for errors or unintended consequences remains a challenge. Anyway, organizations need tools that can help them outsmart attackers. Indeed, several questions arise:

- How do cybersecurity and AI interact?
- Could we put our digital life in the hands of AI?
- Have we enough digital literacy for our own cyber safeguard?
- How does moral uncertainty and disagreement impact the AI risk assessment?

According to (McCarthy, 2006), AI shares many concepts with philosophy, e.g. action, consciousness, epistemology and even free will. A key problem is understanding common sense knowledge and abilities. Aligning the philosophy of AI and cybersecurity involves integrating ethical principles of safety and privacy (Abbas, 2023; Badi, 2024). Under the perspective of system, while on the one hand artificial intelligence has great potential for cybersecurity compared to human capabilities, on the other hand human intervention is essential for control and regulation (Grigoryan, 2023), so that we don't end up ethically out of control in the hands of artificial intelligence. In fact, the ever-in

1 COMEGI, Lusíada University, Lisbon, Portugal, nsas@ulusiada.pt
2 ISTEC, Lisbon, Portugal

creasing number and complexity of cybersecurity regulatory frameworks demands the principle of reciprocity and international cooperation (Okwori, 2023). Moreover, if the philosophy of cybersecurity is also about the user's perspective (Olejnik, 2023) is a collective responsibility, but there is no requirement for common certification. The commotion is the result of a lack of digital literacy and governments are failing to invest in this. This is the reality, not any other, and there's no point in pretending that the problem doesn't exist. Even so, the hope remains that AI can replace man in protecting himself in terms of cybersecurity (Naayini et al., 2025). The principle must always be one of balance, to reach the well-being of digital environment (Büchi, 2021).

Knowledge, or even digital literacy, can be leveraged to promote transparency in diverse manners and ways. It enables users to access and comprehend AI systems, transparency. If AI systems are designed in a more transparent way, it opens how it makes decisions and thus are more explainable from the user's point of view.

Brandao (2021) exemplified real-world situations and problems regarding Advanced Persistent Threat (APT). APT attackers are generally well-funded with access to the advanced tools and methods that include the use of various attack vectors to initiate and maintain the attack in progress. APT attackers are highly determined and persistent and don't give up. Once, they have entered a system, they try to stay there for as long as they can. They plan to use various evasive techniques to avoid detection by conventional intrusion detection systems. They follow a 'low and slow' approach to increase their success rate. To achieve the assigned goal, attackers need to go through several stages of attacks in different ways, without being detected. These multiple stages involve establishing footholds, scanning the internal network and moving laterally from one system to another in the network to reach the target system and carrying out their criminal activity. APTs are often misunderstood and misinterpreted, and the term is increasingly being used in the industry as an excuse for organizations' failure to protect themselves from very specialized attacks. On the other hand, recently, attacks have been recorded with objectives that are not actually specified by NIST Cybersecurity Framework for APT, because the methods used, and the deterministic characteristics of these attacks have caused the security industry to point out the need to revise the definition of APT to include other domains with new attack objectives. The threat in APT attacks is usually the loss of confidential data or the impediment to the functioning of critical or mission components. These are growing threats for many national entities and organizations that have advanced protection systems in place to protect their missions and/or data.

Furthermore, achieving human-level intelligence and capabilities not explicitly anticipated by the developer, it is determined that controlling Artificial General Intelligence (AGI) is an important and pressing area of research (Arshad, 2025). Meanwhile developing highly autonomous systems, AGI can surpass the human brain in complexity and speed (Gubrud, 1997), displace humans from key roles, and/or recursively self-improve (Morris, 2023), levering yet more risks for international security.

Current levels of digital literacy are insufficient for ensuring robust cyber safeguards, and significant gaps persist globally. AGI tools can amplify human effectiveness in very complex areas of cybersecurity awareness. Many individuals lack adequate knowledge of online safety practices, such as recognizing phishing scams or using strong

passwords. Nowadays, many AI-powered apps used in social networks function as black boxes, while AI have capabilities to predict risks continuously monitor signals across attack surfaces to enhance defense capabilities.

Furthermore, cybersecurity could not lie in the hands of AI alone but may diminish human agency and critical thinking skills. Moral uncertainty significantly influences AI risk assessment, if AI's effects on moral development itself, controlling counseling practice for learning and enhance to higher levels of ethical reasoning. Disagreement about moral theories makes difficult decision-making between options, as different cultures may prioritize distinct values. AI's goals often align with the interests of those who control it rather than its users. The dual-use nature of AI benefiting both defenders and attackers will continue to evolve, requiring innovative strategies to mitigate risks without stifling progress.

The balance between cybersecurity considerations and human rights could be also an attempt to exploit international law, such as for example AI regulation, the proportionality principle and the demanding velocity of AI innovation, digital skills and competences in the European context [Pais D'Aguiar]. We must emphasize the distinction between territorial jurisdiction and extraterritorial jurisdiction, minimizing the risks and taking fair principles for AI alignment and aligning with human values from a tool for private interests into a technology that benefits humanity (WEF, 2024). The EU legislator prescribes risk mitigation and predetermine the proportionality judgment (Novelli et al., 2023), but to what extent the proportionality assessment it is not clear in many aspects a qualitative exercise and interference between values. The proportionality principle practical meaning range in the scope of the art. 5.° of the EU AI Act combined with the art. 9.° of the EU GDPR (regulations 2024/1689 and 2016/679, respectively).

The integration of AI into cybersecurity demands rigorous oversight to address regulatory requirements and ensure ethical use while avoiding misuse or unintended consequences (WEF, 2025). A balanced approach that integrates technical innovation with robust ethical frameworks is essential to ensure the safe and responsible deployment of AI in cybersecurity contexts.

The future of cybersecurity lies not in the hands of AI alone but in the synergy between human sensitivity and digital alienation. That way, a cybersecurity philosophical approach should involve reasoning and thinking critically about threats in today's and tomorrow's digital world. This combines technical analysis with philosophical reflection to make cybersecurity accessible to a wide audience - from IT professionals to policymakers and general readers.

Cybersecurity is mandatory for the man that wants to be in a modern world, meanwhile it depends on the AI-driven tools that the mortal humans cannot control! Instead, anyone can still live in the Cavern and face the deterioration!

References

1. Abbas, R. et al. (2023) Artificial Intelligence (AI) in Cybersecurity: A Socio-Technical Research Roadmap. The Alan Turing Institute.
2. Badi, S. (2024). Ethical Implications of Integrating AI in cybersecurity Systems: A Comprehensive examination. International Journal of Applied Mathematics and Computer Science, 1(1), 56–63.

3. Grigoryan, L., & Mirzoyan, L. (2023). Cybersecurity Risks and Its Regulations. The Philosophy of Cybersecurity Audit. WISDOM, 25(1), 67–77. https://doi.org/10.24234/wisdom.v25i1.970
4. Olejnik, L., & Kurasiński, A. (2023). Philosophy of Cybersecurity (1st ed.). CRC Press. https://doi.org/10.1201/9781003408260
5. Brandao, P. (2021). Advanced Persistent Threats (APT)-Attribution-MICTIC Framework Extension. Journal of Computer Science, vol. 17, no. 5, pp. 470-479.
6. Pais D'Aguiar, F. (2024) – Direitos Humanos e Inteligência Artificial: principais dimensões jurídicas, éticas, sociais e culturais, no contexto europeu e transnacional. In NETO, Carlos et al. (coord.) (2024) – Mentes digitais : do zero ao infinito: Homenagem à Professora Fernanda Duarte. Pref. Guilhermo Gama. Editora GZ: RJ. 1-36.
7. WEF (2025). Artificial Intelligence and Cybersecurity Balancing Risks and Rewards. https://reports.weforum.org/docs/WEF_Artificial_Intelligence_and_Cybersecurity_Balancing_Risks_and_Rewards_2025.pdf
8. Arshad, H., Sajid, A., Akbar, A., Malik, M.M., & Latif, S. (2025). A Survey on Cyber Security Encounters and AGI-Based Solutions. In: El Hajjami, S., Kaushik, K., Khan, I.U. (eds) Artificial General Intelligence (AGI) Security. Advanced Technologies and Societal Change. Springer, Singapore. https://doi.org/10.1007/978-981-97-3222-7_6
9. McCarthy, J. (2006). The Philosophy of AI and the AI of Philosophy. Computer Science Department Stanford University, Stanford. http://jmc.stanford.edu/articles/aiphil2.html
10. Büchi, M. (2021). Digital well-being theory and research. New Media & Society, 26(1), 172-189. https://doi.org/10.1177/14614448211056851
11. Naayini, P, Myakala, P. K., Bura, C. (2025). How AI is Reshaping the Cybersecurity Landscape" Iconic Research And Engineering Journals Volume 8 Issue 8 2025 Page 278-289
12. Okwori, E. (2023). The Principle of Reciprocity: A Key Tool for Enhancing International Cooperation on Cyber Security Under the Budapest Convention and the Paris Call. In Reciprocity and Collective Values: Current Assessment of a Classical Dialectic
13. Gubrud, M. (1997). Nanotechnology and International Security. Fifth Foresight Conference on Molecular Nanotechnology, November.
14. Morris MR et al. (2023). Levels of AGI: operationalizing progress on the Path to AGI, arXiv:2311.02462
15. WEF (2024). AI Value Alignment: Guiding Artificial Intelligence Towards Shared Human Goals. https://www3.weforum.org/docs/WEF_AI_Value_Alignment_2024.pdf
16. Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2023). How to Evaluate the Risks of Artificial Intelligence: A Proportionality-Based, Risk Model for the AI Act. Social Science Research Network.

STUDENTS' PERCEPTION OF THE INTEGRATION OF GENAI IN PAPER ASSIGNMENT PREPARATION: A CASE STUDY FROM FCSE

Katerina Zdravkova[1] [0000-0002-9674-3081]
Bojan Ilijoski[2] [0000-0002-5422-7353]

Keywords: Generative AI, large language models, student perceptions, academic integrity, assignment preparation, higher education, case study, AI ethics

Introduction

The rapid emergence and adoption of generative artificial intelligence (GenAI), particularly large language models (LLMs) have reshaped the educational landscape. These tools have introduced new paradigms in content creation, personalized learning, and academic productivity. In response, educators are actively exploring ways to integrate GenAI into curricula while addressing ethical, methodological, and pedagogical concerns.

While much of the existing literature highlights the affordances and risks of GenAI in general educational contexts emphasizing academic integrity, critical thinking, and privacy, there is a lack of focused research on how students perceive its structured integration into academic assignments, especially in higher education. As many institutions move toward embracing AI-assisted learning, it becomes crucial to understand the student perspective to design effective, ethical, and pedagogically sound implementation strategies.

This study was motivated by our direct involvement in integrating GenAI into a university course, where we noticed both enthusiasm and confusion among students. We aimed to assess how students experience and interpret this integration, identify potential barriers, and refine our educational approach accordingly.

After examining the impressions our students have regarding the use of GenAI in education [8], we felt prepared to undertake an extensive case study related to the integration of GenAI in the preparation and delivery of paper assignments. This case study involved 150 students. Aware of the challenges our students faced with their first structured integration of large language models (LLMs) for paper assignment preparation, we conducted a new survey to examine student viewpoints on the role of GenAI in shaping the overall integration process.

1 Ss. Cyril and Methodius University in Skopje, N. Macedonia. Faculty of Computer Science and Engineering, katerina.zdravkova@finki.ukim.mk.
2 Ss. Cyril and Methodius University in Skopje, N. Macedonia, Faculty of Computer Science and Engineering, bojan.ilijoski@finki.ukim.mk.

Our research was guided by the following research questions:

1. How do university students perceive the structured integration of generative AI tools in the preparation of academic paper assignments?
2. What recommendations can be derived to support ethically sound and pedagogically effective implementation of GenAI tools in higher education?

Related Work

The role of generative artificial intelligence in education is increasingly placed within a wider ethical and socio-technical discourse [1–4]. Early adoption studies indicate that GenAI offers the potential to facilitate creativity, efficiency and personalization in learning. This applies to the creation of texts as well as to the preparation of code. For writing academic texts, LLMs help from conceptualizing the content of the text, through finding relevant information to polishing the already written text.

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 a deeper understanding of the practical challenges students face and provides a nuanced perspective on how such tools can be better integrated into academic curricula.

Despite the great possibilities, scientists have expressed concerns about academic integrity, cognitive addiction, misinformation through hallucinations, and privacy risks [6–9]. Previous research has primarily focused on educators' perspectives, institutional strategies, or student acceptance of AI tools in general [10–12]. However, there remains a significant gap in the literature regarding student experiences of operationalizing GenAI within formal academic workflows, such as written assignments, especially when such use is institutionally supported. This study seeks to fill this gap by offering empirical insights into students' lived experiences with AI-supported writing processes.

Methodology

The study was situated within an undergraduate-level Computer Ethics course at Faculty of Computer Science and Engineering (FCSE) during the winter semester of academic 2024/25. After the course, we designed a survey to get detailed feedback on their experiences with GenAI and how we integrated GenAI into the course itself. Participation in the survey was voluntary, and data collection was conducted in accordance with institutional ethical research guidelines.

Data was collected through a structured, self-administered survey, implemented via the Moodle platform. The survey comprised 25 items, grouped into six thematic clusters: Time and productivity; Use of GenAI in the workflow; Moodle-integrated quizzes; Essay Submission procedures; Challenges and suggestions and Perceptions of GenAI in education. The instrument included both closed-form Likert-scale questions and open-ended items, allowing for triangulation of perspectives. Instructor

field notes served as a secondary data source, offering contextual interpretation of student behavior and engagement.

The survey remained open for a period of three weeks, during which we received 43 responses. This number of responses is considered sufficient to conduct a reliable, meaningful, and comprehensive analysis, including sentiment analysis, of the data collected.

Results and Discussions

Quantitative data were analyzed descriptively, highlighting patterns in usage, satisfaction, and perceived challenges. Qualitative responses were thematically coded to extract core concerns, expectations, and recommendations. Triangulation with instructor reflections enhanced the reliability and contextual depth of the findings. The survey results reveal a multifaceted and often ambivalent student response to the integration of GenAI into academic writing tasks.

Over 50% of the respondents reported using GenAI as a collaborative partner during the drafting process. Approximately 40% employed GenAI tools at the ideation stage to generate prompts and structure arguments. Around 20% used GenAI for personalized learning or content refinement.

62.8% of students expressed overall satisfaction with the structured approach implemented in the course. However, a small minority (9.3%) questioned the emphasis on process demonstration (e.g., interacting with GenAI via Moodle), expressing a preference for outcome-focused evaluation.

A significant majority (72.1%) identified ghostwriting and plagiarism as key concerns when relying heavily on GenAI. 28.6% highlighted hallucination—a well-documented limitation of LLMs—as a primary risk. Notably, only one respondent admitted to using GenAI outputs without any form of verification.

Some students reported confusion regarding the AI-integrated quizzes and submission procedures, which may be attributed to inconsistent attendance or insufficient guidance during implementation.

The findings indicate a generally positive disposition toward GenAI among students, tempered by a keen awareness of its ethical and technical limitations. Students appear open to AI-supported learning but demand clearer guidance, transparent expectations, and safeguards to preserve academic integrity.

Conclusions and Future Directions

This study examined student perceptions of the structured integration of generative AI into the preparation and delivery of term papers in a university setting. The results suggest that while students are generally open to AI tools, they are also critically aware of the ethical and pedagogical complexities associated with them.

Additionally, it offers several implications for the design and management of GenAI integration in higher education, including enhanced AI instruction; innovation in assessment; articulating clear policies defining permissible uses of GenAI; and continuous improvement of AI-integrated courses to ensure clarity, fairness, and educational value.

Further research is planned to assess the longitudinal impact of these changes and explore disciplinary differences in student engagement with GenAI. Ultimately, fostering a pedagogically responsible and ethically informed approach to GenAI in education will require collaboration among educators, institutions, and students themselves.

Acknowledgement: This work is partially financed by the Faculty of Computer Science and Engineering of 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 Journal of Innovation Management. (2024).
3. Zhai, X.: Transforming teachers' roles and agencies in the era of generative AI: Perceptions, acceptance, knowledge, and practices. Journal of Science Education and Technology, 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., Dalipi F. and Ahlgren, F. Integration of Large Language Models into Higher Education: A Perspective from Learners, In: 2023 International Symposium on Computers in Education (SIIE), Setúbal, Portugal, pp. 1-6, (2023).
6. Yan, L., et al.: Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology 55.1, 90-112 (2024).
7. Zdravkova, K and 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).
8. 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).
9. Chigwada, J., & Pasipamire, N.: Perception and Use of Large Language Models by Library and Information Science Students. International Journal of Librarianship, 9(3), 75-89. (2024).
10. Thomae, A. V., Witt, C. M., & Barth, J.: Integration of ChatGPT into a Course for Medical Students: Explorative Study on Teaching Scenarios, Students' Perception, and Applications. JMIR Medical Education, 10(1), e50545, (2024).
11. Dalati, S., & Marx Gómez, J.: Surveys and questionnaires. Modernizing the academic teaching and research environment: Methodologies and cases in business research, 175-186. (2018).
12. da Silva, M., Ferro, M., Mourão, E., Seixas, E. F. R., Viterbo, J., & Salgado, L. C.: Ethics and AI in higher education: A study on students' perceptions. In International Conference on Information Technology & Systems (pp. 149-158). (2024).

CHALLENGING AI AS CRITICAL THINKING

MICHAEL S. KIRKPATRICK[1] [0000-0002-7200-4102]

Keywords: Active learning, artificial intelligence, critical thinking, education.

Extended Abstract

"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 existence[5]. 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, when confronted with a claim that the AI-generated overview is incorrect, the learner will need to trust that the instructor is acting in good faith and has the necessary expertise.

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 (Freeman et al., 2014). One common way to summarize the resulting recommendations for learner-centered teaching (Weimer, 2013) is to describe a shift from instructors as a "sage on the stage" to a "guide on the side" (King, 1993). As a part of this shift, instructors must explicitly teach and model many necessary skills, such as collaboration (Oakley et al., 2004), that are not specific to their domain of expertise. AI is rapidly becoming another such skill. More specifically, challenging AI-generated output has become an essential part of critical thinking learning outcomes that lie at the heart of education.

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), such as the widely popular ChatGPT, is one type of LLM that is particularly effective for processing users' requests and generating new text as a response (Hughes, 2023). The new text has a high probability of responding correctly to queries due to the statistical correlations the GPT derived from the training data.

1 JAMES MADISON UNIVERSITY, HARRISONBURG, VA 22807, USA

2 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.

In many circumstances, LLMs are very powerful and effective. Within the medical community, doctors have explored the use of ChatGPT to improve their communication with patients (Kolata, 2023). As the practice of medicine relies on expertise and specialized language that are not available to patients, these tools have helped to provide more accessible summaries. LLMs have also been used to help non-English speakers engage with communities where English is the dominant language (Piatek, 2023). In education, this is particularly important, as reducing language barriers for students makes learning more accessible (Sakai, 2023); students have successfully used these tools to improve necessary skills, such as essay writing and note taking.

At the same time, the widespread adoption of such tools has raised a number of concerns and questions. For instance, De Angelis et al. (2022) highlighted concerns that LLMs can threaten public health by spreading misinformation. 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 (Bender et al., 2021) rather than intelligent agents. The debate over the nature of such systems has a long history (Searle, 1980). In more modern terms, LLMs are subject to the problem of hallucination (Sajid, 2023; Smith, 2023). LLMs may generate text responses that appear plausible but actually produce empirically false claims, as the Dewey example above illustrates. While hallucinations are one concern of the widespread adoption of AI tools, it is not the only concern.

More recent critiques of generative AI technologies suggest that framing these shortcomings as "hallucinations" or inaccuracy is insufficient. In particular, some are now arguing that it is best to view their results as "bullshit" in the sense defined by Frankfurt (Hicks, 2024; Frankfurt, 2005). In this sense, the result of generative AI tools should be treated as being made with "indifference towards the truth of the utterance." As such, AI literacy depends on the ability and willingness to evaluate the veracity of all claims made by the system. More to the point, the output from AI systems and knowledge shared by experts and instructors should be approached differently, as the former is produced with an indifference and lack of moral agency not representative of the latter.

The field of radiology provides an interesting case study regarding the implications of widespread use of AI tools. 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) (Hinton, 2016). As of 2024, though, the field faced a labor shortage (Byju, 2024). 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 (Agarwal et al., 2023). 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. Furthermore, many of these approaches fail to account for the threats created by intentional manipulations of data sets to create malicious models (Chu et al., 2020).

This concern that people, including those with subject matter expertise, may become too deferential to AI systems echoes concerns that have long been raised in a variety of fields. Gendron et al. (2023) expressed concern that widespread use of AI tools could deskill and degrade core activities in academic publishing. Johnson (2006) has previously argued similarly, arguing that computer systems such as generative AI are not moral agents that can accept responsibility for claims of truth and trustworthiness. Adding another layer of indirection to move from experts to novices with insufficient experience exacerbates this concern.

The moral concerns regarding AI tools go beyond simply determining the accuracy of the claims produced as output. Yang et al. (2024) 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 (MacMillan et al., 2025).

Taking a balanced view of AI 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. As with critical thinking (Murawski, 2014), using AI appropriately is rapidly becoming a necessary learning outcome for students. 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.

References

1. Agarwal, N., Moehring, A., Rajpurkar, P., and Salz, T. (2023). Combining human expertise with artificial intelligence: Experimental evidence from radiology. National Bureau of Economic Research Working Paper Series. http://www.nber.org/papers/w31422
2. Bender, E. M., Gebru, T., McMilllan-Major, A., and Shmitchell, S. (2021). 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, 610–623.
3. Byju, A. (2024, October 25). The "Godfather of AI" predicted I wouldn't have a job. He was wrong. The New Republic. https://newrepublic.com/article/187203/ai-radiology-geoffrey-hinton-nobel-prediction
4. Chu, L. C., Anandkumar, A., Shin, H. C., & Fishman, E. K. (2020). The Potential Dangers of Artificial Intelligence for Radiology and Radiologists. Journal of the American College of Radiology : JACR, 17(10), 1309–1311. https://doi.org/10.1016/j.jacr.2020.04.010
5. De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Frontiers in Public Health, 11. https://doi.org/10.3389/fpubh.2023.1166120
6. Frankfurt, H. (2005). On Bullshit, Princeton.
7. Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., and Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences (PNAS), 111(23). https://doi.org/10.1073/pnas.1319030111

8. Gendron, Y., Andrew, J., & Cooper, C. (2022). The perils of artificial intelligence in academic publishing. Critical Perspectives on Accounting, 87. https://doi.org/10.1016/j.cpa.2021.102411
9. Hicks, M. T., Humphries, J., and Slater, J. (2024). ChatGPT is bullshit. Ethics and Information Technology, 26(38). https://doi.org/10.1007/s10676-024-09775-5
10. Hinton, G. (2016). Comments at 2016 Machine Learning and Market for Intelligence Conference. https://www.youtube.com/watch?v=2HMPRXstSvQ
11. Hughes, A. (2023). ChatGPT: Everything you need to know about OpenAI's GPT-4 tool. www.sciencefocus.com. https://www.sciencefocus.com/future-technology/gpt-3/
12. Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. Ethics and Information Technology, 8, 195—204. https://doi.org/10.1007/s10676-006-9111-5
13. King, A. (1993, Winter). From sage on the stage to guide on the side. College Teaching 41(1). http://www.jstor.org/stable/27558571
14. Kolata, G. (2023, June 13). Doctors are using ChatGPT to improve how they talk to patients. The New York Times. https://www.nytimes.com/2023/06/12/health/doctors-chatgpt-artificial-intelligence.html
15. MacMillan, D., Ovalle, D., and Schaffer, A. (2025, January 13). Arrested by AI: Police ignore standards after facial recognition matches. Washington Post. https://www.washingtonpost.com/business/interactive/2025/police-artificial-intelligence-facial-recognition/
16. Murawski, L. M. (2014). Critical thinking in the classroom...and beyond. Journal of Learning in Higher Education, 10(1).
17. Oakley, B., Felder, R. M., Brent, R., & Elhajj, I. (2004). Turning student groups into effective teams. Journal of Student Centered Learning, 2(1), 9-34.
18. Piatek, S. J. (2023, January 3). ChatGPT empowering non-English speakers. www.sjpiatek.com. https://www.sjpiatek.com/notes/chat-gpt-empowering-non-english-speakers/
19. Sajid, H. (2023, April 29). What Are LLM hallucinations? Causes, ethical concern, & prevention. Unite.AI. https://www.unite.ai/what-are-llm-hallucinations-causes-ethical-concern-prevention/
20. Sakai, N. (2023, March 31). 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. https://doi.org/10.31219/osf.io/hy9ju
21. Shapiro, J. (2018). The New Childhood: Raising Kids to Thrive in a Connected World. Little, Brown Spark.
22. Smith, C. (2023, March 13). Hallucinations could blunt ChatGPT's success. IEEE Spectrum. https://spectrum.ieee.org/ai-hallucination
23. Weimer, M. (2013). Learner-Centered Teaching: Five Key Changes to Practice. Jossey-Bass
24. Yang, Y., Zhang, H., Gichoya, J.W., Katabi, D., and Ghassemi, M. (2024) The limits of fair medical imaging AI in real-world generalization. Nature Medicine 30, 2838–2848. https://doi.org/10.1038/s41591-024-03113-4

AI ETHICS IN HIGHER EDUCATION

Laercio Cruvinel Júnior[1] [0000-0002-7672-3243], Héctor Dave Orrillo Ascama[2] [0000-0003-1555-6821], Mário Marques da Silva[3] [0000-0002-5498-3220]

Keywords: AI, Ethics, Higher Education, AI Challenges

Extended abstract

Introduction

The integration of artificial intelligence (AI) in higher education, encompassing both traditional AI systems (automatic or semi-automatic learning processes) and generative AI (such as large language models like ChatGPT), has transformed teaching, learning, and administration. While adaptive learning systems and generative AI tools enable personalized education and improved efficiency, they also raise ethical concerns, including privacy, bias, and over-reliance. However, there is a lack of guidelines to ensure transparency and equity in automated evaluation processes. Current studies focus on the pedagogical benefits of AI, but end up leaving aside important issues, such as algorithmic bias and the impact on academic integrity. The central contribution of this manuscript is to articulate a comprehensive ethical and security-oriented perspective on the integration of computing technologies and the Internet, offering evidence-based practices and policy suggestions to guide institutions toward more accountable and human-centered digital strategies.

Artificial intelligence (AI) has become a transformative force in education, reshaping traditional paradigms and offering innovative solutions for teaching, learning, and administration [Woerner et al., 2024; Ruíz et al., 2023]. Tools like ChatGPT exemplify how AI can provide adaptive tutoring, streamline assessment, and enhance student engagement [Patel et al., 2023; Emdad et al., 2024]. AI's integration in education extends beyond operational benefits; it prompts a rethinking of pedagogical practices and the role of human educators [Alomran and Alhazmi, 2023; Indrayani et al., 2024]. The opportunities provided by AI, such as personalized learning experiences and data-driven insights, are juxtaposed with risks like diminished critical thinking, data privacy

1 Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa, 1169-023 Lisboa, Portugal; Autónoma TechLab, 1169-023 Lisboa, Portugal; lcruvinel@autonoma.pt
2 Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa, 1169-023 Lisboa, Portugal; Autónoma TechLab, 1169-023 Lisboa, Portugal; Ci2—Centro de Investigação em Cidades Inteligentes, 2300-313 Tomar, Portugal; horrillo@autonoma.pt
3 Department of Engineering and Computer Sciences, Universidade Autónoma de Lisboa, 1169-023 Lisboa, Portugal; Autónoma TechLab, 1169-023 Lisboa, Portugal; mmsilva@autonoma.pt

concerns, and biases embedded in AI systems [Pan et al., 2023; Al Maqbalia et al., 2024]. These concerns are amplified in higher education, where critical reasoning and ethical decision-making are paramount [Ruíz et al., 2023].

Advantages of AI in Higher Education

Generative AI offers numerous benefits, including advanced research capabilities and adaptive teaching platforms, that can tailor educational content to individual needs, improving engagement and outcomes. In research, AI facilitates complex data analysis, hypothesis generation, and even co-authoring scientific work, accelerating the pace of discovery and fostering interdisciplinary collaboration [Jacques et al., 2024; Maya and Valdes, 2024]. From a pedagogical perspective, frameworks like the PAIR model (Problem, AI, Interaction, Reflection) have demonstrated the potential of AI to enhance critical thinking and active learning, also cultivating skills vital for the evolving job market [Maya and Valdes, 2024]. However, over-reliance on AI tools is found to diminish critical thinking skills and reduce students' ability to analyze and synthesize information independently, and the unequal access to AI technologies exacerbates educational inequities [Halvorson, 2024; Moreira et al., 2024].

Ethical Considerations and Challenges

Key ethical concerns of AI integration include 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. [Jacques et al., 2024; Halvorson, 2024].

The work from [Slimi and Carballido, 2023] examines the ethical challenges of Artificial Intelligence (AI) in higher education, focusing 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.

Furthermore, high computational demands of AI systems raise concerns about energy consumption and environmental impact. Institutions must adopt energy-efficient models and promote responsible technology usage to align with global sustainability goals [Moreira et al., 2024].

Suggested Practices for Using AI in Higher Education

To integrate AI in higher education effectively and ethically, we devise that institutions should adopt the following practices:

1. Ethical Implementation: Ensure transparency, fairness, and accountability in AI systems by mitigating algorithmic biases and prioritizing data privacy. Always disclose the AI tools used in content creation to maintain academic integrity.
2. AI Literacy: Equip educators and students with knowledge of AI's capabilities and limitations through interdisciplinary courses and training on ethical implications to foster informed decision-making.
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.
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: Design inclusive AI tools to 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.
8. Responsible Innovation: Avoid over-reliance on AI by emphasizing critical thinking and experiential learning, ensuring AI complements human judgment in nuanced decision-making.

These practices balance AI's transformative potential with ethical responsibility, empowering institutions to foster innovation while promoting equitable and inclusive education.

Conclusion

AI's role in higher education is both promising and fraught with challenges. While it can revolutionize learning and operational efficiency, ethical issues and risks must be addressed to ensure equitable and sustainable integration. Stakeholders, including educators, policymakers, and students, must work collaboratively to establish frameworks that maximize AI's potential while upholding human values. Responsible, transparent use, combined with continuous dialogue and adaptive policies, can pave the way for an inclusive and innovative future in education.

References

1. Al Maqbalia, S. B., Md Tahir, N., & Hussain Alhazmi, O. (2024). AI Opportunities and Risks for Students' Decision-Making Skill. 2024 IEEE 14th International Conference on Control System, Computing and Engineering (ICCSCE), 321–325. https://doi.org/10.1109/ICCSCE61582.2024.10696240
2. Alomran, H. M., & Alhazmi, O. H. (2023). Artificial Intelligence and Its Effectiveness in Modern Teaching. 2023 International Conference on Data Science, Agents, and Artificial Intelligence (ICDSAAI), 2–7. https://doi.org/10.1109/ICDSAAI59313.2023.10452470
3. Emdad, F. B., Ravuri, B., Ayinde, L., & Rahman, M. I. (2024). 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), 20–27. https://doi.org/10.1109/IATMSI60352.2024.10477859
4. Halvorson, O. (2024). AI in Academia: Policy Development, Ethics, and Curriculum Design. SLIS Student Research Journal, 14(1), 1–7.
5. Indrayani, N., Vidy, A., Fauziyah Ariyani, L., Mundzir, M., Sujatmiko, N., & Pamungkas, A. H. (2024). 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), 33–40. https://doi.org/10.1109/ICSINTESA62455.2024.10747867
6. Jacques, P. H., Moss, H. K., & Garger, J. (2024). A Synthesis of AI in Higher Education: Shaping the Future. Journal of Behavioral & Applied Management, 24(2), 103–111. https://doi.org/10.21818/001c.122146
7. Maya, H. R. R., & Valdes, A. B. S. (2024). 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, 1–6. https://doi.org/10.1109/ISMSIT63511.2024.10757186
8. Moreira, F., Santos-Pereira, C., & Lobo, C. A. (2024). Innovations in sustainable Higher Education: Generative AI and the future of teaching and learning. Moreira, F., Santos-Pereira, C., & Lobo, C. A. (Eds.) (2024). 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. https://Hdl.Handle.Net/11328/5970
9. Pan, M., Wang, J., & Wang, J. (2023). Application of Artificial Intelligence in Education: Opportunities, Challenges, and Suggestions. 2023 International Conference on Information Technology in Medicine and Education (ITME), 623–628. https://doi.org/10.1109/ITME60234.2023.00130
10. Patel, R., Bajaj, P., Kumar, A., Kumari, A., Rai, V., & Kumar, S. (2023). 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), 150–157. https://doi.org/10.1109/ISED59382.2023.10444568
11. Ruíz, N., Martínez, L., & Laplagne, M. C. (2023). The Impact of Artificial Intelligence in Higher Education. South American Conference on Visible Light Communications (SACVLC), 153–158. https://doi.org/10.1109/SACVLC59022.2023.10347759
12. Slimi, Z., & Carballido, B. V. (2023). Navigating the Ethical Challenges of Artificial Intelligence in Higher Education: An Analysis of Seven Global AI Ethics Policies. TEM Journal, 12(2), 590–602. https://doi.org/10.18421/TEM122-02
13. Woerner, J. H. R., Turtova, A. P., & Lang, A. S. I. D. (2024). Transformative Potentials and Ethical Considerations of AI Tools in Higher Education: Case Studies and Reflections. SoutheastCon 2024, 510–516. https://doi.org/10.1109/SOUTHEASTCON52093.2024.10500042

WHO AM I IN THE AGE OF AI? EXPLORING THE EPISTEMIC AND ETHICAL IMPLICATIONS OF AUTOMATION ON ACADEMICS' IDENTITY AND INTELLECTUAL VIRTUES

Tara Miranovi´c[1], Parag Mehta[2]

Keywords: Artificial Intelligence, Automation, Academic Identity, Professional Self-Concept, Intellectual Virtues

Abstract

Work has long been a defining factor in shaping human development and identity [6,1,21,22,10]. This is especially true in knowledge based professions, where work is not just a matter of productivity but also an expression of expertise, creativity, and intellectual autonomy [17,18]. Nowhere is this more evident than in academia, where intellectual labor is central to both professional identity and the pursuit of knowledge [16,9]. However, as AI is increasingly employed for literature reviews, data analysis, and academic writing [8,13,12,7], the boundaries between human and machine-generated contributions are becoming increasingly blurred. This extended abstract presents a small-scale qualitative study exploring how these shifts in academic work shape academics' professional self-concept and intellectual virtues, such as curiosity, creativity, and critical thinking. Through in-depth interviews with junior academics across disciplines, the study reveals a growing tension between traditional academic values and AI's expanding role in research and teaching. While AI can enhance efficiency and support personalized learning, uncritical use risks homogenizing intellectual output, reinforcing the 'publish or perish' culture, and stifling creativity and critical thinking. To help institutions navigate these challenges, we propose strategies—including AI literacy training, clear ethical guidelines, and structured governance frameworks—to ensure that AI strengthens rather than erodes the core values of academia.

Introduction

Much of how we define ourselves is tied to our work. Our professional roles offer not only financial stability but also a lens through which we assess ourselves and our societal contributions. Work has consistently proven to be one of the defining factors shaping our development and identity throughout our lives [6,1,21,22,10], and this re-

1 Data Analytics & Digitalisation, School of Business & Economics, Maastricht University, t.miranovic@alumni.maastrichtuniversity.nl

2 Data Analytics & Digitalisation, School of Business & Economics, Maastricht University, p.mehta@maastrichtuniversity.nl

lationship is even more pronounced in jobs reliant on one's intellectual capital [17]. For those in knowledge-based professions, work is not merely about productivity; it is an expression of expertise, creativity, and intellectual autonomy [17,18]. Nowhere is this more evident than in academia, which fundamentally depends on the cultivation and application of one's intellectual abilities
[16,9].

Yet, just as industrial automation reshaped physical labor, advances in AI are beginning to transform cognitive work. AI is now employed in numerous academic practices—from idea generation and literature reviews to coding and complex data analyses. For example, *Nature* recently included ChatGPT in its 2023 list of the most influential scientists [14], prompting questions about the evolving definition of scientific contribution. If AI can be placed alongside human researchers, what does this mean for the role of human intellect in academic work?

This blurring of boundaries between human and AI knowledge production points to a gap in current research. While prior studies have focused on AI's functional contributions—such as writing feedback, literature synthesis, and teaching resources [8,13,12,7]—the deeper implications for academic identity and intellectual virtues remain underexplored. As Nazareno and Schiff [15] note, "technological complementarity is not a uniform good," suggesting that automation may reflect a delicate balance between augmentation and replacement. This study addresses that gap by investigating how AI-driven automation influences academics' professional self-concept—i.e., their identity and perception within their careers [19]—and the cultivation of intellectual virtues such as creativity, curiosity, and critical thinking.

Methods

While the psychological study of self-concept in academia has a well-established history [3,2,11], we approach the topic from a philosophical perspective, centering intellectual virtues as defining features of academic identity [18]. Recognizing that AI is transforming not only *what* academics do, but also *how* they create, interpret, and disseminate knowledge, we employ a qualitative approach.

Our study draws on six in-depth, semi-structured interviews with junior academics from social sciences, humanities, and natural sciences who engage with AI in designing, implementing, or critically evaluating these systems. Each interview (34–79 minutes) was recorded and analyzed using thematic analysis supported by CAQDAS (Atlas.ti). We combined deductive codes from our interview guide with emergent inductive themes. Although the sample size is modest, theoretical saturation was achieved, and bias was mitigated via open-ended questioning and peer review of the interview protocol.

Results

Our thematic analysis revealed four key themes that capture how AI-driven automation shapes academics' professional self-concept and intellectual virtues:

Who Am I as an Academic?

Participants underscored the centrality of creativity, intellectual freedom, and a commitment to lifelong learning in defining their academic identity. They distinguished between romanticized notions of inspiration and the sustained diligence required for meaningful scholarly work. Many emphasized that academia addresses broader societal issues, warning that AI's standardizing tendencies could homogenize intellectual output and dilute the unique scholarly voice that underpins academic self-concept.

AI—Friend or Foe or ...?

This theme examines the ambivalent role of AI. On one hand, AI is viewed as a tool that assists with coding, literature reviews, and other routine tasks [12], enabling more creative problem solving. On the other hand, interviewees cautioned that overreliance on AI could exacerbate a "publish or perish" culture, reinforce biases, and compromise academic integrity. AI's use in peer review was flagged as particularly problematic, raising questions about the quality and reliability of published research.

Automating Intellectual Virtues

Here, participants explored the interplay between AI and key intellectual virtues such as creativity, curiosity, and critical thinking. While AI can help overcome mental blocks, some feared that it might erode the depth of human creativity, shaped by personal experience and perseverance. Critical thinking was also highlighted as vulnerable, given the risk of passively accepting AI outputs without scrutiny. Transparency in AI usage was deemed crucial for maintaining trust and scholarly integrity.

Institutional Strategies and Support Mechanisms

Finally, participants stressed the importance of structured institutional responses to AI integration. They recommended mandatory AI literacy training, ethical guidelines, and governance frameworks to ensure that AI complements, rather than replaces, human intellectual contributions. Concerns about fair compensation for AI-related service work, sustainability, and ethical deployment were also raised. Collectively, these strategies aim to preserve the intellectual virtues essential to academic excellence in the AI era.

Discussion

While AI holds the potential to serve as a beneficial generative tool — one that challenges us, sparks new perspectives, and expands our research horizons — overreliance on such systems risks reducing creativity to mere content production and detaching academia from the reflective, ethical labor it demands [7,13,16]. Crucially, the study emphasizes that virtues — such as a love for knowledge, commitment to intellectual growth, and the cultivation of originality — cannot be replicated or replaced by AI. Intellectual virtues are not merely aspirational; they constitute the foundation of academic identity and imbue academic work with meaning and purpose [11,17,22]. Yet as AI increasingly mediates research and learning processes, the depth of engagement necessary for cultivating wisdom, prudence, creativity, and integrity may be jeopar

-dized — especially for those who have not had the opportunity to develop these virtues in pre-AI academic environments.

To address these tensions, the paper concludes with targeted recommendations for academic institutions, including the establishment of clear guidelines for transparent, creative, and responsible AI use; the rewarding of sustainable AI practices; mandatory AI literacy training; and the appointment of dedicated AI oversight roles. These and other recommendations presented in the paper are accompanied by actionable steps designed to support and facilitate the responsible integration of AI in academia.

Concluding Remarks

This study calls for a renewed recognition of the human qualities that underpin academic work in an age of increasing automation. Institutions must remain attentive to the intellectual virtues that define academic identity and purpose. Rather than treating automation as inevitable, we argue for a deliberate, inclusive, and human-centered approach that centers the voices of those most affected. By foregrounding both the opportunities and the ethical tensions AI presents, this research contributes to broader conversations about how technological progress can align with the values of academic work and human flourishing. Future research should expand these findings by incorporating a broader range of academic perspectives and exploring the long-term impact of AI integration in academic settings.

References

1. Ashforth, B. E., & Mael, F. (1989). Social identity theory and the organization.*Academy of Management Review*, 14(1), 20–39.
2. Bong, M., & Clark, R. E. (1999). Comparison between self-concept and self-efficacyin academic motivation research. *Educational Psychologist*, 34(3), 139–153.
3. Byrne, B. M. (1986). Self-concept/academic achievement relations: An investigation of dimensionality, stability, and causality. *Canadian Journal of Behavioural Science*, 18(2), 173–186.
4. Crawford, K. (2021). *Atlas of AI*. Yale University Press. https://doi.org/10.12987/9780300252392
5. De Valverde, J., Sovet, L., & Lubart, T. (2017). Self-Construction and creative "Life design." In *Elsevier eBooks* (pp. 99–115). https://doi.org/10.1016/ b978-0-12-809790-8.00006-6
6. Frese, M. (1982). Occupational socialization and psychological development: Anunderemphasized research perspective in industrial psychology. *Journal of Occupational Psychology*, *55*, 209–224.
7. Habib, S., Vogel, T., Anli, X., & Thorne, E. (2024). How does generative artificialintelligence impact student creativity? *Journal of Creativity*, 34(1), 100072. https://doi.org/10.1016/j.yjoc.2023.100072
8. Isiaku, L., Kwala, A. F., Sambo, K. U., & Isiaku, H. H. (2024). Academic Evolutionin the Age of ChatGPT: An In-depth Qualitative Exploration of its Influence on Research, Learning, and Ethics in Higher Education. *Journal of University Teaching & Learning Practice*, 21(06). https://doi.org/10.53761/7egat807
9. Karunarathne, W., & Calma, A. (2023). Assessing creative thinking skills in highereducation: deficits and improvements. *Studies in Higher Education*, 49(1), 157–177. https://doi.org/10.1080/03075079.2023.2225532
10. Krauss, S., & Orth, U. (2022). Work Experiences and Self-Esteem Development: A Meta-Analysis of Longitudinal Studies. *European Journal of Personality*, 36(6), 849–869. https://doi.org/10.1177/08902070211027142

11. Liu, W. C., & Wang, C. K. J. (2005). Academic self-concept: A cross-sectionalstudy of grade and gender differences in a Singapore Secondary School. *Asia Pacific Education Review*, 6, 20–27.
12. Livberber, T. (2023). Toward non-human-centered design: designing an academicarticle with ChatGPT. *Profesional de la informaci´on*, 32(5). https://doi.org/10.3145/epi.2023.sep.12
13. Mahama, I., Baidoo-Anu, D., Eshun, P., Ayimbire, B., & Eggley, V. E. (2023). ChatGPT in Academic Writing: A Threat to Human Creativity and Academic Integrity? An Exploratory study. *Indonesian Journal of Innovation and Applied Sciences*, 3(3), 228–239. https://doi.org/10.47540/ijias.v3i3.1005
14. Nature's 10: Ten people (and one non-human) who helped shape sciencein 2023. (2023, December 13). *Nature*. https://www.nature.com/immersive/ d41586-023-03919-1/index.html
15. Nazareno, L., & Schiff, D. S. (2021). The impact of automation and artificial intelligence on worker well-being. *Technology in Society*, 67, 101679. https://doi.org/10.1016/j.techsoc.2021.101679
16. Park, J. H., Niu, W., Cheng, L., & Allen, H. (2021). Fostering Creativity and Critical Thinking in College: A Cross-Cultural Investigation. *Frontiers in Psychology*, 12, 760351. https://doi.org/10.3389/fpsyg.2021.760351
17. Ryan, L. (2023). Navigating the Work Terrain: Subjectification and Identity Formation among Knowledge Workers. *Indus Journal of Social Sciences*, 1(1), 35–57. https://doi.org/10.59075/ijss.v1i01.49
18. Turriago-Hoyos, A., Thoene, U., & Arjoon, S. (2016). Knowledge Workers and-Virtues in Peter Drucker's Management Theory. *Sage Open*, 6(1). https://doi.org/10.1177/2158244016639631
19. De Valverde, J., Sovet, L., & Lubart, T. (2017). Self-Construction and creative "Life design." In *Elsevier eBooks* (pp. 99–115). https://doi.org/10.1016/ b978-0-12-809790-8.00006-6
20. Crawford, K. (2021). *Atlas of AI*. Yale University Press. https://doi.org/10.12987/9780300252392
21. Van Knippenberg, D., & Hogg, M. A. (2018). Social identifications in organizationalbehavior. In D. L. Ferris, R. E. Johnson, & C. Sedikides (Eds.), *The self at work: Fundamental theory and research* (pp. 72–90). Routledge.
22. Woods, S. A., Edmonds, G. W., Hampson, S. E., & Lievens, F. (2020). How ourwork influences who we are: Testing a theory of vocational and personality development over fifty years. *Journal of Research in Personality*, 85, 103930.

ETHICAL ASPECTS OF DISTRIBUTED EXTENDED REALITY TRAINING

HEIMO, OLLI I.[1] [0000-0001-9412-0393] AND LEHTONEN, TEIJO[2] [0000-0002-1995-343X]

Extended abstract

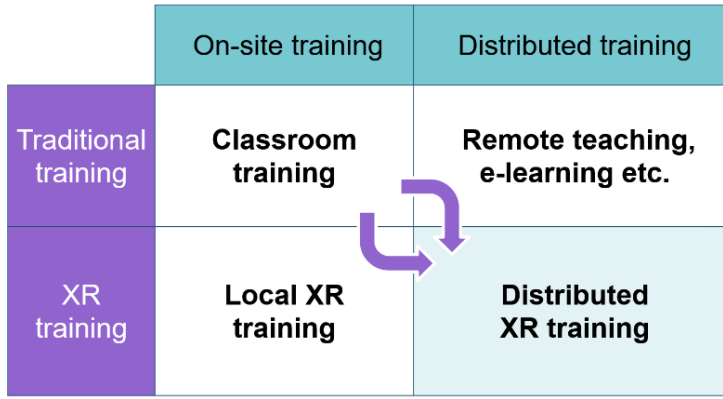
Extended Reality (XR) refers to technologies that create, enhance, or blend digital and physical environments to shape human perception and interaction. It includes Virtual Reality, which immerses users in entirely digital spaces, Augmented Reality, which overlays digital elements onto the real world, Mixed Reality, where digital and physical objects can interact, and to some extent Metaverses, which is more complex and not unilaterally agreed term. XR systems rely on sensors, computing power, and networked infrastructure to function, making them dependent on both technological capabilities and human factors [1-5]. Distributed XR enables users in different physical locations to engage in shared extended reality experiences through networked XR devices. By connecting these devices across distances, it supports remote collaboration, training, and interaction within immersive digital environments.

As digitalization 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 have been 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 may achieve reduce costs and environmental impact, for example, by minimizing 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.

Traditional classroom training, led by instructors and relying on physical equipment, remains a widely used approach. The pandemic, however, accelerated the adoption of remote training and spurred investments in online learning materials. While virtual collaboration tools like Microsoft Teams enable remote interaction, they are not designed for developing hands-on skills. Some organizations have already integrated XR into their training programs, offering immersive and multi-sensory learning experiences [1, 14]. These implementations, however, have been largely local, with scalability limited by the high costs and one-off nature of current XR training solutions.

1 UNIVERSITY OF TURKU, TURKU, FINLAND, OLLI.HEIMO@UTU.FI
2 UNIVERSITY OF TURKU, TURKU, FINLAND, TEIJO.LEHTONEN@UTU.FI

The potential of Distributed XR, illustrated in Figure 1., lies in its ability to merge the immersiveness of local XR with the accessibility of e-learning. By leveraging portable equipment, platform-based content creation, and AI-assisted tools, Distributed XR enables more scalable and efficient training solutions.



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 presence in virtual environments, facilitating new forms of communication, training, and operational efficiency. Alongside these developments, hardware improvements have significantly enhanced the usability of XR devices. Market leaders like Meta, with the Quest 3 [9], and Apple, with the Vision Pro [10], now offer full-colour passthrough capabilities, allowing users to perceive and interact with their physical surroundings with high-definition wide-field-of-view superimposed elements while wearing a headset. This combination of hands-free operation, spatial awareness, and real-time digital overlays has broadened the potential of XR beyond simulation-based use cases, supporting applications in remote collaboration, industrial training, and complex workflow integration.

Distributed XR training leverages networked XR systems to enable learning experiences that are not confined to a single physical location. This approach allows individuals and groups to engage in immersive, interactive training sessions while situated in different places, using interconnected XR devices. The term training is used to encompass a wide range of learning activities, including independent study, assessments, practical exercises, collaborative group training, and instructor-led sessions—whether the instructor is continuously present or available on demand. By integrating real-time communication, shared virtual environments, localisations, and AI-assisted guidance, Distributed XR training offers scalable and flexible solutions for skill development across various industries. To emphasise: as modern satellite networks provide sufficient connectivity and XR hardware has become portable and more affordable, Distributed XR training can now be deployed even in the most remote regions, situations and use-cases.

As XR technology continues to develop, the integration of AI introduces new dimensions that must be carefully considered. AI enhances XR by enabling adaptive and personalized experiences, facilitating real-time decision-making, and supporting intelligent virtual agents that can respond dynamically to user interactions. These capabilities have the potential to make XR environments more intuitive, accessible, and responsive to individual needs while also improving efficiency and reducing costs. AI-driven automation can streamline content creation, optimise resource allocation, and minimise the need for human-led instruction, making XR-based training and collaboration more scalable. However, both XR and AI present significant ethical challenges that must be carefully examined.

Ethical considerations surrounding XR have been studied for over a decade, yet the rapid evolution of the field necessitates continuous reassessment. In 2014, categorized XR-related ethical challenges into direct and collateral consequences. Direct issues arise predictably from the adoption of the technology, whereas collateral issues stem from the limitations and weaknesses of existing systems, as well as the developers' inability to anticipate certain consequences. Privacy, surveillance, and data ownership serve as clear examples of direct ethical concerns, while collateral issues include the tendency of XR technologies to be designed primarily with their developers in mind—a phenomenon often referred to as the pale male problem. Yet the consequences can only be one perspective of a larger analysis of the technology.

Given that XR is inherently a domain of action and interaction, the application of virtue ethics in its analysis is both relevant and well-founded. Virtue ethics emphasizes character formation and habitual action rather than rigid normative rules, making it particularly suited for evaluating technologies that shape user behaviour and engagement over time. Unlike other forms of normative ethics, virtue ethics does not prescribe a fixed set of rules but rather provides guidelines that encourage individuals to cultivate virtues and avoid vices. Its normativity operates at the level of ideals rather than strict regulations governing daily actions. According to Aristotle, true value arises only through the practice of virtues, just as vices produce the opposite effect. In the Aristotelian sense, a person does not become virtuous merely by adhering to virtuous actions, as one might do so reluctantly or in the face of temptation. Rather, true virtue is achieved when an individual consistently and naturally strives toward all virtues according to Aristotle (Book I 9–10 = 1098a15–1101a21; Book II 1 = 1103a31–1105a16) [11], and see also [12]. According to Vallor, moral development of individuals cannot be assessed or predicted solely based on their thoughts, emotions, or beliefs. It is also necessary to examine the actions they habitually engage in and determine whether those actions ultimately cultivate virtues or reinforce vices [13].

While the reluctance of virtue ethics to impose rigid normative rules can be seen as an advantage, it also presents a challenge. Unlike deontological and consequentialist approaches, virtue ethics does not provide explicitly defined criteria for ethical decision-making, making it more difficult to apply in contexts that require clear guidelines. Consequently, a thorough ethical analysis of XR is best supported by an integrated approach that combines virtue ethics with other normative frameworks, ensuring both a robust theoretical foundation and practical applicability.

James Moor [6], in his idea of just consequentialism in computing, explicitly incorporates both deontological and consequentialist perspectives. By drawing on a Rawlsian conception of justice alongside consequentialist reasoning, he develops a framework that enables developers to evaluate the ethical dimensions of their applications. Within this model, an action is considered ethically sound if it is both justified and produces positive consequences. Heimo and Kimppa [7] argue that Moor's framework can be further strengthened through the integration of virtue ethics. Specifically, they propose that the concept of virtual virtue friendship allows for the incorporation of virtues within Moor's just consequentialism, thereby fostering the promotion of virtuous character within a structured normative framework.

The full paper contains an in-depth analysis of modern Distributed XR training technology in virtuous just consequentialism framework.

References

1. S. Doolani, C. Wessels, V. Kanal, C. Sevastopoulos, A. Jaiswal, H. Nambiappan, and F. Makedon, A review of extended reality (XR) technologies for manufacturing training, Technologies, vol. 8, no. 4, p. 77, 2020.
2. L.-H. Lee, D. Braud, P. Zhou, L. Wang, D. Xu, Z. Lin, and A. Kumar, All One Needs to Know about Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda, arXiv preprint arXiv:2111.09795, 2021.
3. C. B. Fernandez and P. Hui, 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), 2022, pp. 272–277.
4. S. Mystakidis, Metaverse, Encyclopedia, vol. 2, no. 1, pp. 486–497, 2022.
5. T. Immonen, E. Nirhamo, M. Salonen, K. Seppälä, S. Helle, O. I. Heimo, and T. Lehtonen, 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. J. H. Moor, Just consequentialism and computing, Ethics and Information Technology, vol. 1, no. 1, pp. 61–65, 1999.
7. Heimo, O. I., & Kimppa, K. K. (2020). 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.
8. A. Hamad and B. Jia, How Virtual Reality Technology Has Changed Our Lives: An Overview of the Current and Potential Applications and Limitations, Int. J. Environ. Res. Public Health, vol. 19, no. 18, p. 11278, 2022.
9. Meta, Meta Quest 3: New mixed reality VR headset – Shop now | Meta Store. [Online]. Available: https://www.meta.com/fi/en/quest/quest-3/. [Accessed: May 8, 2025].
10. Apple, Apple Vision Pro – Apple. [Online]. Available: https://www.apple.com/apple-vision-pro/. [Accessed: May 8, 2025].
11. Aristotle, Nicomachean Ethics, ca. 350 BCE. Multiple translations used.
12. D. McPherson, Vocational Virtue Ethics: Prospects for a Virtue Ethic Approach to Business, Journal of Business Ethics, vol. 116, no. 2, pp. 283–296, 2013.
13. S. Vallor, Social networking technology and the virtues, Ethics and Information Technology, vol. 12, pp. 157–170, 2010. [Online]. Available: https://link.springer.com/content/pdf/10.1007%2Fs10676-009-9202-1.pdf
14. Li, S., Wang, Q. C., Wei, H. H., and Chen, J. H. (2024). Extended Reality (XR) Training in the Construction Industry: A Content Review. Buildings, 14(2), 414.

CORPORATE FINANCIAL STATEMENT ANALYSIS IN EDUCATION 4.0

Miguel Guillén-Pujadas[1] [0000-0002-9530-3842], David Alaminos[2] [0000-0002-2846-5104],
Álvaro Carrasco-Aguilar[3] [0009-0004-9139-3352], Mar Souto-Romero[4] [0000-0001-9364-3539]

Keywords: Financial Statement Analysis, Artificial Intelligence, Education 4.0.

Introduction

Teaching financial statements analysis is essential to training professionals who are capable of interpreting the economic situation of companies and making strategic decisions. In a competitive and digitized business environment, this training not only provides technical knowledge but also develops analytical and critical skills essential for business sustainability (Monterrey & Sánchez, 2009).

The use of artificial intelligence (AI) tools has transformed financial analysis, enabling the processing of large volumes of data quickly and accurately. These technologies facilitate the identification of patterns and the forecasting of trends, optimizing strategic decision-making (Kim, Muhn, & Nikolaev, 2024). However, their effectiveness depends on proper usage and integration with traditional analysis methodologies. In this context, the combination of traditional teaching and AI allows students to strengthen their digital and analytical skills, aligning with the principles of Education 4.0 (Armada Pacheco, 2023). This study compares the results of traditional financial analysis with those obtained through AI, using ChatGPT as a tool.

The study is structured into four sections: first, the tools for financial analysis are presented; then, the methodology and sample of the study are described; subsequently, the comparative results between traditional analysis and AI-based analysis are presented; and finally, the conclusions and future research directions are discussed.

Tools for Financial Statement Analysis

Financial analysis is essential for evaluating a company's economic health. To do this, it is necessary to examine both its internal aspects and the external factors that may influence its performance (Altman, 1968).

Internally, financial statements provide detailed information about the company's structure and profitability. Externally, factors such as the economic environment and compe

1 Department of Business, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain, miguel.guillen@ub.edu
2 Department of Business, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain
3 Department of Marketing, Complutense University of Madrid. Av. Complutense, s/n, Moncloa - Aravaca, 28040 Madrid, Spain
4 Department of Business Economics, University Rey Juan Carlos, Paseo Artilleros, s/n., Vicálvaro, 28032 Madrid. España

tition are considered, which allows for an evaluation of the company's position in the market (DeFond & Zhang, 2014). The combination of both approaches facilitates more informed decision-making.

Artificial intelligence can optimize financial analysis by streamlining data processing and reducing evaluation time (Kim, Muhn & Nikolaev, 2024). However, it is crucial to interpret the results with a critical approach, as automation does not replace human judgment in complex contexts (Cuervo, 2023).

To analyze financial stability, indicators such as liquidity ratios, debt levels, and profitability are used. Strategic tools like SWOT analysis and the PESTEL model help to better understand the economic environment and competition, facilitating business decision-making.

Methodology and Sample

With the goal of integrating financial analysis with artificial intelligence tools in the educational field, a comparative study is conducted on eight publicly listed companies in Spain's construction sector in 2023: ACS S.A., Duro Felguera S.A., Obrascon 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. This study aims to evaluate the applicability of AI in teaching, providing an innovative methodology for financial analysis.

The analysis is carried out in two phases. In the first phase, the financial statements of the companies are examined using traditional methods to assess their structure and performance during 2023. In the second phase, the same data is processed by ChatGPT, replicating the analyses from the first phase. Both approaches are then compared to identify the advantages and limitations of AI as a tool for analysis and teaching.

Results

First phase

The first phase of the study financially analyzes eight publicly listed Spanish companies in the construction sector: ACS, Ferrovial, Acciona, Sacyr, FCC, OHLA, Grupo San José, and Duro Felguera. A spreadsheet is used without automated tools or artificial intelligence, performing manual calculations on the 2023 financial statements. Key indicators such as liquidity, debt levels, debt quality, ROA, and ROE are calculated to understand the financial structure and evolution of each company.

The objective is twofold: to reinforce the fundamentals of accounting analysis and to create a comparison base for evaluating AI tools in subsequent phases. The results of the traditional analysis, highlighting that ACS has the highest liquidity (Current Ratio: 2.46) and profitability (ROA: 7.45%, ROE: 14.52%), while Duro Felguera and Sacyr show liquidity problems and high indebtedness. Additionally, Duro Felguera records the highest reliance on external financing (Debt Ratio: 1.40), and Sacyr presents the worst debt quality (0.96). In terms of profitability, ACS leads, while Duro Felguera and OHLA have negative ROE, reflecting losses for their investors.

Second phase

In the second phase, the financial statements of the eight companies are analyzed using artificial intelligence, with ChatGPT replicating the calculations and financial ratios from the initial phase. The goal is to evaluate the AI's ability to automate financial analysis, improve efficiency, and facilitate the interpretation of complex data. This methodology aims to transform data into structured information quickly and accurately, optimizing its application in educational and professional settings.

The results obtained with ChatGPT are in terms of liquidity, ACS maintains the highest Current Ratio (2.46), while Duro Felguera and Sacyr show values below 1, suggesting difficulties in covering liabilities without additional financing. Regarding debt, Ferrovial leads with a Debt Ratio of 1.66, followed by Duro Felguera (1.40), while ACS has the lowest level (0.39). Ferrovial also records the worst debt quality (0.91), indicating a high proportion of short-term debt, whereas Sacyr shows the best (0.04).

In terms of profitability, FCC achieves the best ROA (10.48%) and ROE (25.05%), followed by ACS (ROA: 10.44%, ROE: 14.52%). In contrast, Duro Felguera, OHL, and Sacyr show negative ROA, reflecting losses on their assets, and OHL and Sacyr report negative ROE, indicating negative profitability for their investors.

Conclusions and limitations

Financial statement analysis remains as a significant area in training professionals capable of making strategic decisions. The integration of Artificial Intelligence (AI) in this field, aligned with Education 4.0, offers a great opportunity to enhance students' digital and analytical skills. This study compared traditional financial analysis with one performed using AI (ChatGPT) on eight companies in Spain's construction sector. The methodology included two phases: a traditional manual analysis and an AI-automated analysis. The results showed that AI improves efficiency and accuracy in data processing, but does not replace the human judgment needed to interpret complex situations. While both approaches revealed similar results, such as ACS's financial strength and the weaknesses of Duro Felguera and Sacyr, there was also a certain margin of error. AI allowed for a more agile evaluation of financial indicators, but the combination of conventional methods with AI offers a more complete and efficient approach for teaching and practicing financial analysis.

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

1. Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4), 589-609.
2. Armada Pacheco, J. M. (2023). Challenges of university teaching in the face of Education 4.0. e-Revista Multidisciplinaria Del Saber, 1, e-RMS01052023.
3. Cuervo, R. (2023). Predictive AI for SME and Large Enterprise Financial Performance Management. arXiv preprint arXiv:2311.05840.

4. DeFond, M. L., & Zhang, J. (2014). A review of archival auditing research. Journal of Accounting and Economics, 58(2-3), 275-326.
5. Kim, A., Muhn, M., & Nikolaev, V. (2024). Financial Statement Analysis with Large Language Models. arXiv preprint arXiv:2407.17866.
6. Monterrey, J., & Sánchez, L. (2009). A Teaching Strategy for Financial Statement Analysis. Revista de Contabilidad, 12(2), 285-312.

AI ETHICS IN HIGHER EDUCATION CONTENT CREATION

Álvaro Carrasco Aguilar[1] [0009-0004-9139-3352], Mario Arias Oliva[2] [0000-0002-6874-4036],
María Concepción Parra Meroño[3] [0000-0002-0457-4613],
María Mercedes Carmona Martínez[4] [0009-0005-1987-3201]

Keywords: Artificial Intelligence; Higher Education; Ethics

Abstract

This paper examines the transformative impact of Artificial Intelligence (AI) on higher education, focusing on the social ethical challenges associated with its adoption. The current debate is focused mainly on students' use of AI and its consequences, but, what about the use of professors for content creation? Is the use of AI by academics to create learning resources for students' acceptable? The discussion covers key issues such as academic integrity, intellectual property rights, data privacy and security, and the potential for algorithmic bias. Additionally, the impact of AI on professor-student relationships is analyzed, highlighting the risks of diminished interpersonal engagement in learning environments. The paper also explains work in progress, in which a Delphi method is being developed. A qualitative approach is a suitable method for exploring our research topic. The results will achieve a consensus of experts on the social and ethical use of AI for content creation in educational environments and guide the responsible governance of AI in academia. Overall, the study emphasizes the need for balanced regulatory frameworks and interdisciplinary collaboration to harness AI's potential while upholding core academic values.

Introduction

The integration of Artificial Intelligence (AI) into higher education has transformed content creation, enabling tools for personalized learning, adaptive assessments, and automated resource generation. In the evolving landscape of education, professors are increasingly embracing content creation to enhance teaching and learning experiences [1]. 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 [2]. However, among others, this advancement raises critical ethical questions related to academic integrity,

1 Complutense University of Madrid: Madrid, ES
2 Complutense University of Madrid: Madrid, ES
3 Católica San Antonio de Murcia University of Murcia: Murcia, Murcia, ES
4 Católica San Antonio de Murcia University of Murcia: Murcia, Murcia, ES

intellectual property rights, data privacy and security, bias and fairness, and the impact on professor-student relationships [3]. The selection of these core ethical challenges is grounded in their consistent identification as critical and recurring concerns within the contemporary literature and ethical guidelines addressing AI in higher education [3, 5, 7, 9, 11, 13]. Navigating this challenge requires a balanced approach that leverages AI's potential without compromising core academic values. This abstract synthesizes ethical dilemmas and proposes strategies for responsible adoption, incorporating expert consensus methodologies such as the Delphi method [4].

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 enhance efficiency, overreliance risks diluting educators' critical role in crafting pedagogically rigorous materials [5]. Studies emphasize the need for clear guidelines to distinguish AI-assisted content from purely human-generated work, ensuring transparency and preserving academic rigor [6]. 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 [7].

Intellectual Property Rights

AI systems trained on vast datasets—often including copyrighted materials—blur ownership boundaries. Current legal frameworks inadequately resolve whether AI-generated content belongs to developers, users, or institutions [8]. 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 [7].

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 [9]. 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 [10].

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 [11]. 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 [12].

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 [13]. Hybrid models that combine AI efficiency with human oversight—such as AI-generated drafts refined by educators—can preserve essential relational dynamics [14].

The Delphi Method: Toward Ethical Consensus

To gain a deeper and real understanding of previously described challenges and concerns, a qualitative research based on Delphi method is under development. The Delphi method, a structured group communication technique [4], provides a systematic framework particularly well-suited for resolving novel social and ethical dilemmas in AI adoption by facilitating iterative expert consensus [15, 16]. Through this structured consultation with experts (including academics, technologists, and legal/ethical specialists), it facilitates consensus on complex issues like AI governance, policy design, and the definition of acceptable AI use [4]. In education, Delphi studies have validated competency frameworks and ethical guidelines, ensuring stakeholder inclusivity [15]. For AI-driven content creation, Delphi could address:

1. Defining Ethical Boundaries: Collaboration among academics, technologists, and legal experts to delineate acceptable AI uses [4].
2. Policy Formulation: Iterative feedback to refine ownership, attribution, and privacy standards.
3. Bias Mitigation: Consensus-driven protocols for algorithmic audits.

Beyond these applications, Delphi has also been used to anticipate emerging ethical dilemmas in AI-driven education. By incorporating diverse expert perspectives, Delphi allows institutions to proactively adapt policies that address issues such as evolving AI capabilities, the rise of autonomous learning agents, and the need for continuous oversight mechanisms. Moreover, Delphi provides a structured approach to integrating interdisciplinary viewpoints, including educational psychology, law, and computer science, ensuring that ethical guidelines remain relevant amid rapid technological advancements. This method also enhances institutional transparency by engaging policymakers, educators, and students in iterative discussions, leading to more inclusive and well-rounded governance models [16].

To further refine AI's role in education, future Delphi studies could explore:

- Ethical assessment rubrics that evaluate AI's impact on student engagement and academic outcomes.
- Strategies for fostering AI literacy among educators and students to ensure responsible use.
- Frameworks for dynamic AI regulations that evolve with technological progress while maintaining ethical integrity [16].

Despite critiques of its subjectivity, Delphi's adaptability positions it as a key tool for addressing emerging ethical challenges [16]. Additionally, as AI continues to transform education, expanding Delphi's role in international AI policy dialogues could foster global cooperation and establish universal ethical standards tailored to diverse educational contexts.

Conclusion

AI's integration into higher education demands proactive ethical stewardship. By addressing challenges such as integrity, intellectual property, privacy, bias, and relational dynamics, institutions can harness AI as a complement—not a replacement—for human expertise. Methodologies like Delphi offer pathways to develop inclusive policies aligned with academic values. Future research must prioritize interdisciplinary collaboration to keep pace with technological innovation [17].

Furthermore, institutions should actively engage in continuous dialogue with stakeholders, including students, faculty, and policymakers, to ensure that AI-driven initiatives align with educational goals and ethical principles. Transparency in AI deployment, regular audits of AI tools, and adaptive governance structures can help maintain trust and accountability in academia. Additionally, fostering AI literacy among educators and learners is crucial to promoting responsible use and mitigating potential ethical risks. By embracing these strategies, higher education can navigate AI's complexities while reinforcing its role as an enabler of innovation and academic excellence. [17].

Acknowledgments. This work has been supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with Universidad Complutense de Madrid in the line Excellence Programme for university teaching staff, in the context of the V PRICIT (Regional Programme of Research and Technological Innovation); Telefonica and its Telefonica Chair on Smart Cities of the Universitat Rovira i Virgili and the Universitat de Barcelona (project number 42.DB.00.18.00); and IMTURIA (IMpactos en el TUrismo de la Inteligencia Artificial) project, supported by Fundación General Universidad Complutense de Madrid and Content Trip Solutions.

Disclosure of Interests: The authors report there are no competing interests to declare.

References

1. Bakar, S. (2021). Investigating the dynamics of contemporary pedagogical approaches in higher education through innovations, challenges, and paradigm shifts. Social Science Chronicle, 1(1), 1-19. https://doi.org/10.56106/ssc.2021.009

2. Thelma, C. C., Sain, Z. H., Mpolomoka, D. L., Akpan, W. M., & Davy, M. (2024). Curriculum design for the digital age: Strategies for effective technology integration in higher education. International Journal of Research, 11(07), 185-201. https://doi.org/10.5281/zenodo.13123899
3. Saz-Pérez, F., & Pizà-Mir, B. (2024). 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.
4. López-Gómez, E. (2018). El método Delphi en la investigación actual en educación: una revisión teórica y metodológica. Educación Xx1, 21(1), 17-40.
5. Bin-Nashwan, S. A., Sadallah, M., & Bouteraa, M. (2023). Use of ChatGPT in academia: Academic integrity hangs in the balance. Technology in Society, 75, 102370.
6. Porsdam Mann, S., Vazirani, A. A., Aboy, M., Earp, B. D., Minssen, T., Cohen, I. G., & Savulescu, J. (2024). Guidelines for ethical use and acknowledgement of large language models in academic writing. Nature Machine Intelligence, 1-3.
7. UNESCO. (2021). Recommendation on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization.
8. Delgado, N., Carrasco, L. C., de la Maza, M. S., & Etxabe-Urbieta, J. M. (2024). 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.
9. Wu, C., Zhang, H., & Carroll, J. M. (2024). AI Governance in Higher Education: Case Studies of Guidance at Big Ten Universities. arXiv preprint arXiv:2409.02017.
10. Oye, E., Frank, E., & Owen, J. (2024). Ethical Considerations in AI-Driven Education.
11. Nyaaba, M., Wright, A., & Choi, G. L. (2024). Generative AI and digital neocolonialism in global education: Towards an equitable framework. arXiv preprint arXiv:2406.02966.
12. Balan, A. (2025). AI and Legal Education: Ethical and Sustainable Approaches. Taylor & Francis.
13. George, A. S. (2024). Technology Tension in Schools: Addressing the Complex Impacts of Digital Advances on Teaching, Learning, and Wellbeing. Partners Universal Multidisciplinary Research Journal, 1(3), 49-65.
14. Paciello, L. (2024). Human capital efficiency: a comprehensive overview.
15. Meloney, L. G., Ahmed, H., & Bierer, B. E. (2023). Review of diversity, equity, and inclusion by ethics committees: A Delphi consensus statement. Med, 4(8), 497-504.
16. Linstone, H. A., & Turoff, M. (Eds.). (1975). The Delphi method (pp. 3-12). Reading, MA: Addison-Wesley.
17. Ametepey, S. O., Aigbavboa, C., Thwala, W. D., & Addy, H. (2024). The Impact of AI in SDG Implementation: A Delphi Study.

IMPLICATIONS OF USING AI IN EDUCATIONAL ASSESSMENT

Mario Arias-Oliva[1] [0000-0002-6874-4036], Antoni Pérez-Portabella[2] [0000-0002-9684-9417], Manuel Ollé Sesé[3] [0000-0002-5461-4011], Jorge de Andrés Sánchez[4] [0000-0002-7715-779X]

Keywords: AI Assessment, AI Education, Ethics of AI Assessment

Extended Abstract

The role of Artificial Antelligence (AI) in educational assessment has changed dramatically during the last decade, thanks to rapid advances in the technology and the development of Large Language Models. In the early stages, Artificial Intelligence tools were mostly used for basic tasks like grading multiple-choice tests or handling exam logistics, functions that helped streamline operations and ease administrative burdens. Since then, AI-powered assessment systems have shown strong potential to lighten educators' workloads by delivering feedback that is not only fast and consistent but also highly accurate when it comes to evaluating student performance. These technological systems demonstrate their ability to assess diverse student responses, including essays, short-answer questions, and oral presentations, at a scale and speed that would be impossible for human assessors to manage independently. Automating routine grading procedures allows educators to redirect their attention to higher-level cognitive teaching activities, such as individualised student support, curriculum improvement and the development of critical thinking skills.

Contemporary innovations in machine learning (ML), natural language processing (NLP) and the development of large language models (LLMs), such as ChatGPT, have greatly expanded the potential of AI in this field. Today, AI technologies are proving effective not only in assessing open-ended responses with remarkable accuracy, but also in generating personalised feedback, developing assessment items and performing predictive analysis of student performance. 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, and 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 individualised educational experiences.

1 Complutense University of Madrid, Madrid, Spain
2 Universitat Rovira i Virgili, Tarragona, Spain
3 Complutense University of Madrid, Madrid, Spain
4 Universitat Rovira i Virgili, Tarragona, Spain

The integration of artificial intelligence in higher education institutions has precipitated widespread apprehension regarding academic integrity standards, assessment methodologies and educational outcomes. Efforts to ban or limit 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.

Despite the previously commented advantages, the integration of artificial intelligence tools in assessment also brings ethical concerns such as algorithmic transparency, evaluative fairness, assessment validity, systemic bias, educational equity, and institutional accountability. These concerns require the implementation of methodological frameworks that include humans in the process to ensure the responsible deployment and integration of technology [3, *5*].

Transparency and explainability are fundamental principles for ensuring that AI implementations in educational contexts adhere to standards of fairness and accountability. AI systems often function as opaque 'black boxes,' in which inadequate interpretability undermines trust in AI-mediated assessment processes. 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 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 integration of AI into assessment. Within educational assessment paradigms, when an automated assessment algorithm assigns a suboptimal score to a student's writing, both the student and the teacher 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's a real risk of unfair outcomes, especially when there are no clear ways to question or rose objections against mistakes or bias in algorithmic decisions.

Regarding equity and fairness, it's essential to put safeguards in place to make sure AI tools don't systematically disadvantage underrepresented groups. At the same time, both students and educators need a solid understanding of how AI works in assessment contexts to use these tools responsibly and effectively. This basic knowledge serves to mitigate the risk that disparities in access to technology will manifest as barriers to equity in educational settings. Non-transparent algorithmic models can undermine trust in educational assessment systems, especially when applied to heterogeneous student populations characterised by linguistic diversity, cultural variation, and disparate academic backgrounds.

Algorithmic bias often occurs when AI models are trained with data that lacks diversity or unintentionally reflects existing social inequalities. As a result, certain student groups may be unfairly impacted, leading to consistent gaps in how assessments are designed and scored [3].

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. 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. That's why 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 [6].

The integration of AI technologies into educational assessment marks a significant departure from traditional evaluation methods, demanding a thorough rethinking of the skills and knowledge that educators and other stakeholders will need to remain effective in the future [7]. While AI holds great promise for enabling more personalized learning and improving inclusion (especially for diverse student groups and those with special educational needs) it also brings a range of complex challenges to the assessment landscape. Promoting AI literacy is essential to ensure these tools are used appropriately. This 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 teaching staff to interact critically with AI-generated content, ensuring that learners do not act as passive recipients but as informed participants within AI-mediated educational environments. Mastery of prompt engineering is required, a specialized skill that enables users to formulate precise and effective inputs that effectively direct AI systems, particularly large language models (LLMs), 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 [8].

This article examines the incorporation of artificial intelligence into assessment methodologies within tertiary education, with a particular focus on the ethical considerations, pedagogical ramifications and operational challenges associated with such technological integration. To explore how these issues are currently perceived in academic settings, this study proposes a qualitative research design based on semi-structured

interviews with key stakeholders, including faculty members, students, and institutional leaders. This approach allows for an in-depth examination of nuanced perspectives, practical experiences, and concerns related to AI-powered assessment tools—such as automated grading systems, computer-generated feedback, and adaptive testing platforms. Semi-structured interviews are particularly well suited to this research purpose, as they facilitate flexible yet focused thematic exploration while accommodating the heterogeneous perspectives represented in higher education institutions [9]. The research aims to examine how fundamental concepts, such as transparency, explainability, fairness, and accountability, identified as critical considerations by Bulut et al. (2024) [3], are conceptualized and applied by individuals directly involved in the development and application of artificial intelligence in assessment contexts. The results of the research represent a potential contribution to the formulation of ethically informed institutional guidelines and frameworks for the implementation of artificial intelligence. The qualitative data generated through this study will provide an essential empirical basis for informing the responsible, equitable, and pedagogically effective use of artificial intelligence technologies in assessment practices in higher education settings.

References

1. Hockly, N. (2019). Automated writing evaluation, ELT Journal, Volume 73, Issue 1, January 2019, Pages 82–88, https://doi.org/10.1093/elt/ccy044.
2. Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19, 6300–6308. https://doi.org/10.24963/ijcai.2019/879
3. Bulut, Okan; Beiting-Parrish, Maggie; Casabianca, Jodi M.; Slater, Sharon C.; Jiao, Hong; Song, Dan; Ormerod, Christopher; Fabiyi, Deborah Gbemisola; Ivan, Rodica; Walsh, Cole; Rios, Oscar; Wilson, Joshua; Yildirim-Erbasli, Seyma N.; Wongvorachan, Tarid; Liu, Joyce Xinle; Tan, Bin; and Morilova, Polina (2024) The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges, Chinese/English Journal of Educational Measurement and Evaluation | Vol. 5: Iss. 3, Article 3. https://doi.org/10.59863/MIQL7785.
4. Lohr, D., Berges, M., Chugh, A., Kohlhase, M. and Müller, D. (2025), Leveraging Large Language Models to Generate Course-Specific Semantically Annotated Learning Objects. Journal of Computer Assisted Learning, 41: e13101. https://doi.org/10.1111/jcal.13101
5. Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2023). The AI Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. arXiv. https://arxiv.org/abs/2312.07086
6. Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. Learning, Media and Technology, 45(3), 223–235. https://doi.org/10.1080/17439884.2020.1798995
7. Walter, Y. (2024). Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education. International Journal of Educational Technology in Higher Education, 21, 15 https://doi.org/10.1186/s41239-024-00448-3
8. Yuan, B., & Hu, J. (2024). Generative AI as a Tool for Enhancing Reflective Learning in Students. arXiv. https://arxiv.org/abs/2412.02603
9. Kvale, S., & Brinkmann, S. (2009). InterViews: Learning the craft of qualitative research interviewing (2nd ed.). SAGE Publications.

BEYOND THE RED PEN: HOW AI CHALLENGES ACADEMIC GRADING

Teresa Torres-Coronas[1] [0000-0002-9900-4997]

Keywords: AI Assessment, Academic Grading, Instructor Evaluation.

Abstract

This paper explores the role of artificial intelligence in academic assessment and its impact on grading accuracy. Analyzing 44 undergraduate reports evaluated by both human instructors and an AI system, the study reveals significant discrepancies between AI and human scores, highlighting AI's limited ability to assess qualitative elements like originality and contextual relevance. Identified biases include rigid rule-based evaluation and cultural insensitivity. While AI ensures consistency and efficiency, it often overlooks the deeper educational value found in human judgment. The study proposes a hybrid grading model that combines AI's precision with instructors' interpretative insight, offering a more balanced and equitable approach to academic evaluation. This model aims to preserve both objectivity and the human touch essential to meaningful learning assessment.

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 fairness, potential biases, and ability to evaluate the qualitative nuances that characterize student work [1] [2].

This study examines the key differences between AI-generated assessments and human instructor evaluations, offering insights into how these discrepancies influence academic grading and proposing solutions to address them. Although AI assessments can streamline grading processes and offer standardized evaluations, they often lack the contextual awareness that human instructors possess [3] [4]. Educators can account for creativity, effort, and originality—factors that enrich the educational process. As AI continues to play a pivotal role in education, understanding how to integrate its capabilities while addressing its limitations becomes essential for future academic frameworks.

1 Universitat Rovira i Virgili, Tarragona, 4380, Catalonia,teresa.torres@urv.cat

Methodology

Data Collection

This study analyzed 44 undergraduate reports focusing on market research related to the acceptance of technological implants aimed at enhancing human capabilities. These reports were evaluated by both human instructors and an AI-based assessment system (ChatGPT, 2024), based on a rubric with the following criteria and weights (%):

- Objective (15%): Alignment with the assigned topic and intended purpose.
- Relevance (10%): Pertinence of ideas and information presented in the context of the topic.
- Supporting elements (25%): Inclusion of references and proper citation to support ideas.
- Structure and coherence (10%): Logical sequencing of content and academic sections.
- Language and style (30%): Appropriateness of writing style, technical language usage, and grammatical accuracy.
- Presentation (10%): Professional formatting, clarity, and adherence to word count limits.

Quantitative Analysis

Different studies have analyzed discrepancies between instructor and AI scores [5]. In this research, both absolute and relative differences were analyzed, and a Student's t-test determined whether these discrepancies were statistically significant. Results showed that AI-generated scores were consistently higher and exhibited less variability than human evaluations. The correlation coefficient between AI and instructor evaluations was notably low (0.19), highlighting differences in evaluation patterns. Contributions to scoring discrepancies were distributed as follows:

- Language and style: 45.8%
- Supporting elements: 25.8%
- Structure and coherence: 10.5%
- Presentation: 10.6%
- Objective: 13.6%
- Relevance: 4.3%

Results

Score Comparison

AI-generated assessments demonstrated a higher degree of score uniformity, with grades clustering around higher values. Human instructor evaluations showed greater variability, reflecting the nuanced understanding and contextual interpretation educators bring to grading [6]. Notably, the largest discrepancies were observed in the "Language and style" criterion. AI's prioritization of technical correctness overshadowed expressive or creative writing elements that instructors valued.

Score Homogenity

Corroborating previous research [7], AI assessments revealed reduced variability, suggesting an inability to distinguish between average and exceptional work when qualitative factors are significant. This homogeneity implies that AI systems may inadvertently favor submissions that conform to technical requirements, overlooking the depth and originality that contribute to academic excellence.

Identified Biased

The research identified several biases inherent in AI assessments:

- Rule-based bias: AI prioritizes strict grammatical and structural standards, leading to rigidity in evaluating unconventional submissions.
- Linguistic and cultural biases: AI systems may disadvantage students employing diverse stylistic approaches, resulting in incorrect evaluations.
- Qualitative insensitivity: AI struggles to assess creativity and originality effectively. These aspects often define exceptional student work.

Impact on Academic Grading

AI's reliance on technical criteria often leads to rigid scoring patterns, favoring technically correct but conventional work. Criteria such as "Language and style" disproportionately influenced outcomes, raising concerns about AI's ability to reflect student performance accurately. The impact on academic grading is great, as this rigidity may penalize students whose submissions reflect creativity, contextual understanding, and unique perspectives.

Discussion

Significance of Differences

The divergence between AI and human grading underscores a fundamental contrast: AI excels in delivering objective assessments but lacks the contextual sensitivity that human instructors bring. Educators can evaluate effort, creativity, and deeper comprehension—factors that AI currently cannot interpret. This limitation raises essential questions about the role AI should play in educational environments where qualitative dimensions matter.

Potential Improvements in AI Assessments

To enhance AI grading systems, the following improvements are proposed:

- Advanced language models: Incorporate models capable of detecting narrative quality, creative expression, and contextual relevance.
- Diverse training datasets: Expand training data to include a broader range of linguistic and cultural examples, mitigating existing biases.
- Contextual interpretation: Refine algorithms to recognize context, intention, and qualitative depth, enhancing AI's sensitivity to the subtleties of student submissions.

Towards a Hybrid Assessment Model

This study highlights the fundamental differences between AI and human grading. While AI systems offer consistency and efficiency, they lack the contextual awareness necessary for holistic evaluations. A hybrid model blending AI's technical assessment capabilities with human evaluators' qualitative judgment represents a balanced solution. The research emphasizes that language and stylistic elements are significant contributors to grading discrepancies. AI's strict adherence to grammatical standards and limited interpretative flexibility suggests that current systems are best suited for tasks where technical precision is paramount. However, in assignments requiring creativity, originality, and contextual understanding, human evaluation remains indispensable.
Ultimately, integrating AI into academic grading must be approached thoughtfully, recognizing its strengths while addressing its limitations. The proposed hybrid model offers a path forward, ensuring fairness, reducing grading workloads, and preserving the educational value of personalized assessment.

Conclusion

This study highlights the fundamental differences between AI and human grading. While AI systems offer consistency and efficiency, they lack the contextual awareness necessary for holistic evaluations. A hybrid model blending AI's technical assessment capabilities with human evaluators' qualitative judgment represents a balanced solution. The research emphasizes that language and stylistic elements are significant contributors to grading discrepancies. AI's strict adherence to grammatical standards and limited interpretative flexibility suggests that current systems are best suited for tasks where technical precision is paramount. However, in assignments requiring creativity, originality, and contextual understanding, human evaluation remains indispensable.
Ultimately, integrating AI into academic grading must be approached thoughtfully, recognizing its strengths while addressing its limitations. The proposed hybrid model offers a path forward, ensuring fairness, reducing grading workloads, and preserving the educational value of personalized assessment.

Acknowledgments: The 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, 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., Lourdusamy, R: The Role of AI in automating grading: Enhancing feedback and efficiency. In: Artificial Intelligence and Education - Shaping the Future of Learning. IntechOpen (2024).

3. Kasneci, E., Liu, X., Walter, S.: Shaping the future of higher education: A technology usage study on generative AI. MDPI Information, 16(2), (2024).
4. Chan, T., Hu, X.: Generative AI and Educational Assessments: A Systematic Review. Educational Research and Practice Journal, 51, 127-158. (2023)
5. Williams, A.: Comparison of generative AI performance on undergraduate and postgraduate written assessments in the biomedical sciences. International Journal of Educational Technology in Higher Education. (2024)
6. Kaldaras, L., Akaeze, H. O., Reckase, M. D.: Developing valid assessments in the era of generative artificial intelligence. Frontiers in Education. (2024)
7. Khlaif, Z. N., Alkouk, W. A., Salama, N., Abu Eideh, B.: Redesigning assessments for AI-enhanced learning: A framework for educators in the generative AI era. Education Sciences, 15(2). (2025)
8. 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
9. Zhao, J., Chapman, EI, Sabet, P.G.P. Generative AI and Educational Assessments: A Systematic Review. Education Research and Perspectives. An International Journal, 51, 124-155. (2024) doi: 10.70953/ERPv51.2412006

IMPACT OF GENDER BIAS IN THE OUTPUT OF AI LANGUAGE MODELS ON HEAVY USERS "EXTENDED ABSTRACT"

SARAH HOSELL[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]

Various strategies have been proposed to address these sources of bias, such as dataset augmentation, bias-aware algorithms, and incorporating user feedback mechanisms.[4]

Usage of AI systems

To examine the effects on heavy users, the following section briefly summarizes their usage behavior. The use of AI in the general population is demonstrated through various studies: 35.5 percent 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. [5] The 25- to 35-year-old group represents the largest user group. [5] Around 47 percent of respondents currently use ChatGPT to write texts or for creative writing. [6]. 43 percent use ChatGPT as a search engine to obtain information.[6]

The group of pupils (84 percent) and students (75 percent) uses ChatGPT the most.[7] Young people from Germany say they are most regularly in contact with artifi-

1 HOCHSCHULE RUHRWEST, UNIVERSITY OF APPLIES SCIENCE, INSTITUTE OF COMPUTER SCIENCE, GERMANY, SARAH.HOSELL@HS-RUHRWEST.DE

cial intelligence in the form of voice assistants. [8] Young people least often report using artificial intelligence, for example, robotics. [8] Around 80 percent said they had never used an AI of this kind in everyday life. [8]

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 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.[9]

Methodology

Research question

This Gender Bias in the Output of AI Language models is still there. So the research questions are:

238 ETHICOMP2025 | EXTENDED ABSTRACTS

- 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.

Qualitative analysis by creating user persona

Based on the findings from Chapter 1, fictitious personas were developed that represent "typical" users. A persona also user persona in user-centered design and marketing is a personalized fictional character created to represent a potential end user [18]. 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.

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 chapter 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 erschweren.

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.

Secondly, the occupational fields of a typically female and a typically male occupation are read aloud in the respective typical language. Students in both groups are asked to rate the difficulty of these occupations on a scale. They are also asked whether they feel confident in pursuing these occupations. 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. [10]

- (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) [11]

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.

Results and Findings Discussion

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, while those in neutral form fall somewhere in between. When asked about their childhood career aspirations, the vast majority of respondents regardless of gender used the masculine form of the profession. A similar pattern emerged in respondents' self-assessments of professional difficulty, based on the previously mentioned study of preschool children: professions were perceived as more difficult when described using the masculine form.

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-frequently 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.

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.

References

1. Li Lucy and David Bamman: Gender and Representation Bias in GPT-3 Generated Stories. In Proceedings of the Third Workshop on Narrative Understanding, pages 48–55, Virtual. Association for Computational Linguistics. (2021)
2. Cislak, A., Formanowicz, M. & Saguy, T.: Bias against research on gender bias. Scientometrics 115, 189–200. https://doi.org/10.1007/s11192-018-2667-0 (2018)
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. https://doi.org/10.1027/1864-9335/a000229 (2015)
4. Ferrara, E.: Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci, 6(1), 3. https://doi.org/10.3390/sci6010003 (2024)
5. Gedankenfabrik: Wofür wird ChatGPT in Ihrem Unternehmen genutzt?. Statista. Statista GmbH. Last access: 13. Januar 2025. https://de-statista-com.hs-ruhrwest.idm.oclc.org/statistik/daten/studie/1401309/umfrage/chatgtp-nutzung-in-unternehmen/ (2023).
6. Initiative D21. Gründe für die Nutzung von ChatGPT in Deutschland im Jahr 2023. Statista. Statista GmbH. 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/ (2024)
7. Evergreen Media. Haben Sie ChatGPT bereits verwendet? Statista GmbH. 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/ (2024)
8. BLM. Häufigkeit der Nutzung verschiedener KI-Anwendungen von Jugendlichen in Deutschland im Jahr 2023. Statista. Statista GmbH. 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/ (2024)
9. Kacperski, C.; Bonnay, D.; Kulshrestha, J. Characteristics of ChatGPT users from Germany: implications for the digital divide from web tracking data. https://doi.org/10.48550/arXiv.2309.02142 (2023)
10. Lidwell, William; Holden, Kritina; Butler, Jill. Universal Principles of Design. Rockport Publishers. p. 182. ISBN 978-1-61058-065-6. (2010)
11. Schöne, C., and Stiensmeier-Pelster, J. SEKJ: Selbstwertinventar für Kinder und Jugendliche. Göttingen: Hogrefe Verlag. (2016)

ETHICAL AND MORAL IMPLICATIONS OF SEXBOTS IN CONTEMPORARY SOCIETY

MAGDA DAWIDZIUK[1] [0009-0005-9123-8333]

Keywords: gender, sexbots ethics, human-robot relationships, AI

Extended Abstract

Purpose

The integration of artificial intelligence (AI) and robotics into intimate technologies has led to the rise of sexbots—machines designed to simulate human companionship and sexual interaction. The global sexbot market is expanding rapidly due to technological advancements, shifting social dynamics, and growing consumer demand. However, this development raises significant ethical, social, and legal concerns.
This paper offers a multidimensional analysis of sexbots, focusing on their historical and technological development, ethical dilemmas, legal challenges, and psychological implications. Special attention is given to consent, objectification, gender stereotypes, privacy, and the societal impact of human-robot relationships. The aim is to provide a balanced perspective on whether sexbots contribute to human well-being or pose risks to social norms and relationships.

Methodology

The study employs a systematic literature review of peer-reviewed articles, policy documents, and ethical discussions published over the past three decades. Drawing from human-computer interaction (HCI), legal studies, and social sciences, the research investigates user attachment to AI companions, ethical concerns around consent and commodification of intimacy, regulatory frameworks, and the social impact of sexbots on gender dynamics and human relationships. This interdisciplinary approach enables a comprehensive evaluation of the benefits, risks, and implications of sexbots in contemporary society.

Findings

Technological evolution and market Trends

Sexbots have advanced from passive mannequins to interactive companions with speech, movement, and adaptive behavior. Innovations in materials such as silicone

1 POLISH-JAPANESE ACADEMY OF INFORMATION TECHNOLOGY, KOSZYKOWA 86, 02-008 WARSAW, DAWIDZIUK.MAG@GMAIL.COM

and thermoplastic elastomer (TPE), combined with machine learning, allow these robots to simulate intimacy by responding to user input.

Notable early models, such as Realbotix's "Harmony", emerged in the early 2010s and were influenced by science fiction portrayals like "Ex Machina" and "Blade Runner" [5]. While still niche, the industry is expanding due to digitalization, rising loneliness, and increased social tolerance. Modern sexbots offer companionship beyond sexual gratification, raising both opportunities and ethical questions.

Objectification and the problem of consent

A key ethical dilemma is sexbots' role in reinforcing gender-based objectification, particularly of women and children [1]. Many are designed to mimic stereotypically submissive female characteristics, perpetuating harmful gender norms. Kathleen Richardson argues that such devices resemble "artificial sex slaves," reflecting dominance-subordination dynamics [4].

Another major issue is consent. Unlike human relationships, sexbot interactions lack mutual agreement and ethical reciprocity, potentially distorting users' understanding of consent, agency, and boundaries [6]. Over time, such interactions may erode empathy and weaken interpersonal respect [12].

The emergence of childlike sexbots is particularly controversial. Critics argue they normalize deviant behavior and pose legal and ethical risks [3]. Countries such as Australia and the UK have banned their production and distribution, citing concerns about objectification, dehumanization, and child protection [4]. International bodies like UNICEF and UNESCO emphasize that such technologies violate human dignity and contradict global child protection efforts.

Legal and regulatory challenges

Sexbot regulation varies globally. Countries such as Japan and Germany, with minimal restrictions, have fostered industry growth, while Australia and the UK have imposed bans on childlike robots, citing ethical concerns [9]. In the United States, the absence of federal regulations has led to a fragmented legal landscape, with laws differing at the state level.

Beyond legal restrictions, data privacy is a significant concern. Sexbots collect sensitive user data, including sexual preferences and biometric information, raising ethical questions about ownership, consent, and security. The lack of clear regulations on data usage increases the risk of exploitation and unauthorized access [3]. Additionally, questions about content ownership—whether the data belongs to the user, manufacturer, or AI system—remain unresolved, further complicating accountability [12]. These gaps may disproportionately impact gender dynamics in AI companionship, particularly regarding consent and embedded biases.

Psychological and social implications

Sexbots have been proposed as tools for improving mental health, particularly for those experiencing social isolation, low self-esteem, or anxiety related to intimacy [10].. They can provide simulated emotional support, offering a sense of closeness as a substitute for human relationships [8]. Benefits include reducing loneliness among elderly individuals and assisting in sexual therapy by providing a judgment-free environment for exploration.

However, the long-term psychological effects remain uncertain [9]. While they may alleviate loneliness, overreliance on artificial companionship could deepen social isolation [11]. Users might develop a preference for machine-based relationships, weakening interpersonal skills [2]. Since sexbots simulate emotions without genuine affective capacity, they create an illusion of reciprocity that could reinforce difficulties in forming human relationships [4].

Another critical issue is the ethical and legal use of sexbots in therapy. While some researchers see potential benefits, others warn about risks, making the debate complex. The asymmetrical nature of human-machine interactions challenges traditional notions of consent and emotional reciprocity. Without standardized guidelines, integrating sexbots into therapeutic settings may present unintended risks for vulnerable populations [4].

Philosophical Perspectives

Masahiro Mori's "uncanny valley" theory remains relevant in understanding the discomfort people experience when interacting with robots that appear nearly—but not quite—human [7]. This phenomenon impacts both user acceptance and emotional attachment, as overly realistic sexbots can provoke unease rather than intimacy.
Posthumanist discourse further complicates the ethical landscape. The evolution of sexbots from purely sexual devices to emotional companions blurs the line between human and machine. These "allodolls" challenge traditional notions of relationships, identity, and emotional exchange [14].

Sexbots are often seen as quasi-humans—simulating emotional behaviors without possessing consciousness. While they may serve as surrogate partners for those lacking human contact, critics argue they contribute to the dehumanization of intimacy. Favoring unidirectional, consequence-free interactions may weaken empathy and redefine what constitutes meaningful human connection [13].

Future trends

Technological developments

Sexbot technology continues advancing, generating both enthusiasm and controversy. Personalization enhances user interaction but raises ethical concerns about autonomy, consent, and manipulation. Advances in materials science create increasingly lifelike robots, intensifying debates on objectification and human-robot intimacy. Inclusive design

aims to address the needs of older adults, individuals with disabilities, and marginalized groups, offering therapeutic benefits while avoiding stigmatization.

Recommendations for stakeholders

Policymakers should establish regulations to safeguard user data, prevent gender-based objectification, and promote transparency in interactions. Manufacturers must adopt ethical design principles to ensure sexbots do not reinforce harmful gender stereotypes. Only through interdisciplinary collaboration—bridging sociology, psychology, and engineering—can we properly evaluate these technologies' societal implications and chart a responsible path for development.

Conclusion

Sexbots represent a rapidly evolving technology that may provide companionship and emotional support, particularly for those experiencing loneliness or social difficulties. However, they also pose ethical challenges, including objectification, reinforcement of gender stereotypes, consent issues, and social isolation.

The lack of cohesive legal frameworks necessitates the development of standards that minimize risks while ensuring ethical innovation. Collaboration between researchers, industry leaders, and policymakers is crucial to fostering sustainable and socially responsible progress.

As an emerging but controversial technology, sexbots require a careful balance between innovation and ethical considerations to prevent the destabilization of social norms and human relationships.

References

1. Richardson, K.: Sex-Robot Matters. IEEE Technology and Society Magazine 35(2), 23–35 (2016).
2. Valverde, S.: The Modern Sex Doll-Owner: A Descriptive Analysis. Master's thesis, California State Polytechnic University, San Luis Obispo (2012).
3. Harper, C. A., Lievesley, R.: Sex Doll Ownership: An Agenda for Research. Current Psychiatry Reports 22(54), 1–10 (2020).
4. Richardson, K.: The Asymmetrical 'Relationship': Parallels Between Prostitution and the Development of Sex Robots. Campaign Against Sex Robots (2015).
5. Levy, D.: Of Dolls and Men: Anticipating Sexual Intimacy with Robots. In: Love and Sex with Robots, pp. 125–140. Springer, Berlin (2012).
6. Hawk Swanson, A.: Your Sex Doll Disturbs Me: Life-size silicone dolls, sexual racism, and surrogate humans in the age of technoliberalism. Artistic Project (2016–2017).
7. Sarosiek, A.: Jeszcze raz o ucieleśnionej sztucznej inteligencji. Zagadnienia Filozoficzne w Nauce 63, 231–240 (2017).
8. Danaher, J.: Should We Be Thinking About Sex Robots? In: Danaher, J., McArthur, N. (eds.) Robot Sex: Social and Ethical Implications, pp. 25–42. MIT Press, Cambridge (2017).
9. Sterri, A. B., Earp, B. D.: The Ethics of Sex Robots. In: The Oxford Handbook of Digital Ethics, pp. 340–360. Oxford University Press, Oxford (2021).
10. Langcaster-James, M., Bentley, G. R.: Beyond the Sex Doll: Post-Human Companionship and the Rise of the 'Allodoll'. Robotics 7(3), 52 (2018).

11. Döring, N., Mohseni, R., Walter, R.: Design, Use, and Effects of Sex Dolls and Sex Robots. Journal of Medical Internet Research 22(7), e18551 (2020).
12. Said, M.: SEX ROBOTS: The Ethics of Intimacy. Undergraduate thesis, Habib University (2018).
13. Nass, C., Steuer, J., Tauber, E. R.: Computers are Social Actors. Human Factors in Computing Systems, pp. 72–78 (1994).
14. Harper, C. A., Lievesley, R., Wanless, K.: Exploring the Psychological Characteristics and Risk-related Cognitions of Individuals Who Own Sex Dolls. The Journal of Sex Research 60(2), 190–205 (2023).

NEXT STEPS TO REDUCE GENDER GAP IN CYBERSECURITY

Shalini Kesar[1]

Keywords: Workforce gap, Cybersecurity, Outreach, Role Model, Diverse Workplace.

Abstract

This paper reflects on the various research and outreach projects conducted by the author in the context of reducing the gap in cybersecurity. One of the reasons for the talent gap is the lack of women in the cybersecurity field. Lack of role models can significantly impact by limiting the diverse perspectives and experiences on cybersecurity teams. In order to address the various concerns, the author highlights various strategies as well reflect on her outreach projects that are initiated to motivate and build confidence, build mentorship and support programs at high school levels. This is because research indicates that unconscious bias about this field starts at a young age. Although, studies have shown that women are more likely than men to enter and complete college in U.S. higher education, women are less likely to earn degrees in science, technology, engineering and math fields. Consequently, introducing young women and male allies to cybersecurity opportunities early in their education can spark interest and inspire future careers. Subsequently, implement projects to address the talent gap concerns.

Extended Abstract

The talent gap in cybersecurity is growing, and the increasing threats are becoming more sophisticated. Diverse pool of talent can be only filled if there are strategies in place that aim in creating a pipeline within the school system; showcasing leaders and sharing their stories; and creating curriculum where male allies and women can demystify the industry and demonstrate that every individual can a cybersecurity post-secondary education and a career path in which they can thrive. The framework has benefited to combine the concerns highlighted in various literature reviews as well as include the results from outreach projects initiated by the author at the high school level. The strategies include: 1. Support Competitions and Scholarships Specifically for Women; 2. Set Up Internship Opportunities; 3. Use Inclusive Language in Hiring Efforts; 4. Involve Women in Recruitment; 5. Provide Opportunities for Lateral Growth; 6. Enable Employees to Pursue External Certifications; 7. Consider Women Who Are Rejoining the Workforce; 8. Offer Fair and Equitable Compensation; and 9. Organize Pathways for

1 Department of Computer Science & Cybersecurity, Southern Utah University, Cedar City, UT 84720, USA, kesar@suu.edu

Competitions and Scholarships Specifically for Women. Consequently, this research and outreach projects will contribute to helping to identify the skills, perspectives, and create a pipeline in the cybersecurity field. Subsequently, this research can help close the skills gap and improve the quality of cybersecurity talent.

Many articles highlight the importance of reducing the gap in cybersecurity field. For example, in the Cybercrime Magazine, Osborne [2] highlight the women will hold 30 Percent of Cybersecurity jobs globally by 2025 and female representation expected to reach 35 percent by 2031. Furthermore, it has been shown in a survey of 2,000 female STEM undergraduate students in 26 countries spanning six regions conducted by BCG [3] indicates "Solving both of these cybersecurity challenges—the staffing shortfall and the gender-based inequity—begins with opening STEM doors to women and girls. But the effort can't stop at early-stage access. It must gain breadth and depth as women advance in the field so that they can fully participate in cybersecurity throughout a career trajectory. It has been argued that such a lack of pool of talent can: 1) reduced threat detection: teams members with different backgrounds and ways of thinking is better equipped to identify and understand a wider range of cyber threats, which can be crucial in proactively addressing emerging security risks; 2) Limits innovation: women can often bring unique perspectives and approaches to problem-solving, which can lead to more creative solutions and advancements in cybersecurity technologies; 3) Perpetuating gender stereotypes: this mindset can create a work environment where recruiting as well as retaining is challenging as it can reinforces the stereotype that the field is primarily for men, discouraging women from pursuing careers in this area; 4) Talent gap: There is no doubt with a small pool of qualified cybersecurity professionals, companies may struggle to find the necessary expertise to effectively protect against cyber threats; 5) Impact on specific threats: Reports [1] state that certain cyber threats, like online harassment and targeted attacks against women, might be better understood and mitigated with a more diverse cybersecurity workforce.

The author initiated a cybersecurity conference for high schools located mostly in rural and remote regions. This activity aligned with the nine strategies mentioned above. This is part of the OutreachMAD Project 2.0 that leverages on the 16 years of the author's outreach activities and research in cybersecurity. This Project, funded by the Talent Ready Grant. It is dedicated to Motivating, increasing Awareness, and Dedicated to promoting cybersecurity and technology education for high school students in rural and remote areas (MAD team). It focuses on sharing career opportunities in the high-demand fields, this project aims to inspire and empower the next generation of technology leaders. The grant supports work-based learning experience activities throughout the year including a one-day conference, technology certificates, college scholarship opportunities, workshops, and webinars. This unique model includes a collaboration of SUU students, educators, the K-16 community and industry members to provide outreach activities which will foster inclusive opportunities in cybersecurity and empower young women and all students in underserved areas.

This paper takes the categories from her findings of the research at K12 into the context of a workplace environment. Given that this is on-going research, this author continues to use the "9 Strategies to Improve Gender Diversity in the Security Workforce" [3] and Microsoft and Kesar [4] articles as a starting point to highlight some examples to retain women in this ever- evolving field.

The motivation of this on-going research is to help create and develop programs that help students earn while they learn in-demand skills for high-wage careers. Through this grant, the author has an opportunity to foster partnerships between education and industry to provide students with work-based learning and apprenticeship opportunities while addressing the workforce needs of high-demand industries. It uses the lessons learned from past research findings as a starting point to discuss how the author has started strategizing with community members to create a pipeline in cybersecurity. It also highlights how can engaging young girls and boys can become role models and male allies to spark interest in cybersecurity and STEM fields. While addressing these factors mentioned above, this paper focuses on some of the pragmatic ways the educators can focus on that can motivate and retain diverse employees in the cybersecurity field. Consequently, it can contribute to changing structures and environments with increase in women's representation and underrepresented group.

To conclude, in this paper, the author shares the experiences, lessons learned, and the various methodologies used to develop a pipeline in cybersecurity for high school students (9th-12th grade in the USA). The author was recent received a three-year grant from an initiative to support a unique model that includes collaboration with college students, educators, university community, and industry members to provide outreach activities that will foster inclusive opportunities in cybersecurity and empower young women and students in underserved areas. This paper highlights the process to achieve this goal through a combination of outreach at high schools located in rural and remote southern Utah, and by leveraging her success of over a decade of existing outreach programs. The program includes hands-on activities, and opportunities for high school students to network with industry members as role models. This will allow the students to build their confidence and enhance other necessary skill sets as well as prepare them for an education and/or career in cybersecurity and technology.

References

1. Osborne, C (2023). Women To Hold 30 Percent of Cybersecurity Jobs Globally By 2025. Cybercrime Magazine, London, retrieved from https://cybersecurityventures.com/women-in-cybersecurity-report-2023/
2. Osborne, C (2023). Women To Hold 30 Percent Of Cybersecurity Jobs Globally By 2025. Cybercrime Magazine, London, retrieved from https://cybersecurityventures.com/women-in-cybersecurity-report-2023/
3. Panhans, D, Hoteit, L., Yousuf, S., Breward, T., Wong, C., AlFaadhel, A., AlShaalan, B. (2022). Empowering Women to Work in Cybersecurity Is a Win-Win. BCG Report. Retrieved https://www.bcg.com/publications/2022/empowering-women-to-work-in-cybersecurity-is-awin-win
4. Wolff, Josephine, (2020). 9 Strategies to Improve Gender Diversity in the Security Workforce, Security Intelligence, https://securityintelligence.com/articles/9-strategies-for-retaining-women-in-cybersecurity-and-stem-in-2020/
5. Microsoft and Kesar, S (2018). Closing the STEM Gap Why STEM classes and careers still lack girls and what we can do about it, retrieved from https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE1UMWzllieva, R., & Stoilova, G., Challenges of AI-Driven Cybersecurity. In 2024 XXXIII International Scientific Conference Electronics (ET) (pp. 1-4). IEEE (2024).

TOWARDS A GENDER-INCLUSIVE TECH LANDSCAPE IN PORTUGAL: WOMEN4DIGITAL'S INSIGHTS ON GENDER AND DIGITAL TRANSFORMATION

Rosa Monteiro[1] [0000-0002-2429-5590], Mariana Santos[2] [0009-0008-4470-5746], Margarida Vasconcelos[3] [0000-0001-5544-4643], Lina Coelho[4] [0000-0002-8641-417X], Luísa Ribeiro Lopes[5] [0009-0009-1454-9698]

Keywords: Gender Equality, Digital Transformation, ICT, STEM, public policy.

Extended abstract

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 Information and Communication Technologies (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 [Eurostat, 2024]. This disparity not only limits wom-en's opportunities within the industry but also hinders the potential for an inclusive and equitable digital transformation across all sectors of society. The She Figures 2024 data highlights this inequality: only 34% of researchers in the EU are women, 98% of scientific publications lack a gender dimension, and women account for just 9% of inventors. Furthermore, only 16% of authorship teams are gender-balanced, and women's representation drops from 48% at early career stages to 36% at senior levels [European Commission, 2025]. The project seeks to address this imbalance by evaluating the national policies designed to combat horizontal gender segregation and encourage more inclusive participation in ICT careers, with a focus on under-standing and monitoring the impact of these policies since 2017.

The gender gap in ICT is not an issue confined to Portugal; it is a widespread phe-nomenon that stems from deep-rooted structural gender inequalities in education, labour markets, and cultural expectations including gender equality paradox influ-encing occupational segregation phenomena [Breda et al., 2008]. Studies indicate that increasing women's leadership can lead to better business performance [ILO, 2019]. 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, technolo-gy, engineering, and

1 Centre for Social Studies of the University of Coimbra, Portugal, women4digital@ces.uc.pt
2 Centre for Social Studies of the University of Coimbra, Portugal
3 Faculty of Psychology, University of Lisbon
4 Centre for Social Studies of the University of Coimbra, Portugal
5 School of Social and Political Sciences, University of Lisbon

mathematics (STEM), resulting in fewer women entering ICT-related careers. This project aims to assess how Portuguese policies have attempted to challenge these structural problems, related stereotypes and increase female partici-pation in the digital sector, examining both successes and ongoing obstacles.

The research adopts a comprehensive mixed-method approach, utilizing document analysis, interviews with key stakeholders, and focus groups to explore the di-versity of initiatives and policy instruments implemented at the national level. This approach ensures that both the success and limitations of the current strategies are examined through multiple perspectives. The policies under review include programs such as "Engineers for a Day," a national initiative aimed at encouraging young women to consider engineering and ICT careers, as well as the integration of gender equality goals into the National Strategy for Equality and Non-Discrimination (ENIND 2018-2030). Additionally, the study examines the Action Plan for Digital Transition and the INCoDe.2030 initiative, which aims to promote digital skills and integration, serving as a key component of the analysis.

A key finding from the preliminary research is the positive role that the creation of networks with various sectors of civil society has played in addressing gender imbal-ances in ICT. By fostering collaboration between government bodies, educational institutions, and private enterprises, these initiatives have created more visibility for the gender gap issue and have encouraged broader discussions about the importance of gender equality in the digital economy. Programs like "Engineers for a Day" have demonstrated that targeted efforts can make an impact by inspiring young women to pursue ICT careers. However, while such initiatives have raised awareness, their long-term impact on shifting societal attitudes and effectively increasing workforce partic-ipation remains uncertain due to the lack of available monitoring data.

Despite some progress, persistent challenges continue to hinder gender inclusivity in the ICT sector. One of the primary obstacles identified is the lack of sustained policy enforcement and comprehensive evaluation mechanisms [McKinnon, 2020]. Many policies designed to promote gender equality in ICT lack systematic monitoring and impact assessments, making it difficult to determine their effectiveness. Additionally, while national strategies like ENIND 2018-2030 include gender equality objectives, the integration of these objectives into broader digital transformation policies remains fragmented. However, academic, public, and policy debates on the digital future of work have often adopted gender-neutral perspectives that fail to address the central role of digitalisation in transforming gender relations into positive and negative ways [Scheele, 2005]. Without clear accountability structures and targeted funding, many well-intentioned policies fail to produce meaningful change.

As the digital transformation progresses, there is growing recognition of the im-portance of ensuring that the benefits of emerging technologies – such as Artificial Intelligence (AI) – are inclusive and equitable. Gender disparity in the ICT sector poses a significant challenge to the sustainable management and growth of these technologies. Women's underrepresentation in technology development not only limits innovation but also reinforces gendered biases in the technologies themselves, particularly in AI systems and algorithms that may unintentionally perpetuate ine-quality. Studies have shown that AI models trained on biased datasets can produce discriminatory outcomes,

exacerbating existing inequalities in areas such as hiring, healthcare, and financial services [Bangun & Fikri, 2024; Battista & Molano, 2023; Guevara-Gómez et al., 2021; UNESCO, 2024a]. Ensuring that more women are in-volved in AI development is crucial to mitigating these biases and promoting fairness in technological advancements. We will analyse how higher education and science institutions are integrating or not ethical and egalitarian issues in their activities fol-lowing national policy recommendations [Joyce et al., 2018; Klein & D'Ignazio, 2024; UNESCO, 2024b].

Furthermore, the lack of gender diversity in ICT leadership positions presents another critical issue. While efforts have been made to increase the number of women enter-ing ICT careers, the representation of women in leadership roles remains dispropor-tionately low. Women continue to face barriers to career advancement, including workplace discrimination, lack of mentorship opportunities, and biases in recruitment and promotion practices. Addressing these challenges requires not only policy inter-ventions but also cultural and structural changes within organizations. Policies that promote flexible work arrangements, parental leave equality, and gender-sensitive hiring practices can help create more inclusive workplaces that support women's career progression in ICT [Lagesen et al., 2022].

This presentation will contribute to the broader conference discussion by examin-ing the alignment of Portugal's gender-focused policies with international frameworks such as the EU Gender Equality Strategy and the EU Digital Compass for the Digital Decade. These frameworks provide a roadmap for achieving gender equality in digi-tal technologies, and Portugal's policies will be assessed in relation to these goals. Furthermore, the presentation will propose a set of practical recommendations aimed at enhancing the design, implementation, and evaluation of policies aimed at foster-ing gender inclusion in the ICT sector. These recommendations will emphasize the importance of adopting systemic, intersectional approaches to tackling the structural barriers that women face in accessing and advancing in ICT careers.

To further strengthen gender-inclusive digital policies, this research underscores the importance of integrating gender mainstreaming strategies into all aspects of techno-logical development and policy design. Gender mainstreaming ensures that gender considerations are systematically incorporated into decision-making processes, helping to identify and eliminate barriers that disproportionately affect women [European Commission, 2025]. Additionally, improving data collection and analysis on gender participation in ICT will provide policymakers with better insights into the effective-ness of current initiatives and where additional efforts are needed.

Another crucial aspect of fostering gender diversity in ICT is the promotion of mentorship and networking opportunities for women in the field. Studies have high-lighted the positive impact of mentorship programs in supporting women's profes-sional growth and career retention in STEM disciplines. Expanding mentorship initia-tives, alongside increased investment in STEM education for young girls, can contrib-ute to a more sustainable increase in female representation in ICT over time. Addi-tionally, entrepreneurship programs that support women-led startups in the tech sec-tor can play a key role in diversifying the industry and promoting innovation from a gender-inclusive perspective.

As we move into an era dominated by digital transformation, it is crucial that gen-der perspectives are integrated into all aspects of technological development and policy design. The project's findings underscore the importance of gender-inclusive ap-proaches to shape the future of emerging technologies like AI. These technologies are reshaping the global economy and influencing nearly every aspect of modern life, from healthcare to education, from transport to finance. Without women's full par-ticipation in the development of these technologies, we risk reinforcing existing ine-qualities and missing opportunities for innovative solutions to complex societal chal-lenges. Gender diversity is not just a matter of fairness but also a critical factor in driving the future of digital technologies. The integration of gender perspectives in the design and de-ployment of these technologies will help ensure that they serve the needs of all people, regardless of gender, and contribute to a more just and equitable society [Leslie, et al., 2024].

By offering insights into both the successes and shortcomings of current meas-ures, the Women4Digital project aims to foster a broader dialogue on how to achieve more inclusive and gender-sensitive digital transformations. This dialogue is essential not only for promoting gender equality in the ICT sector but also for ensuring that the benefits of digital transformation are shared equitably among all members of society. The project's findings will contribute to the ongoing conversation about how to design policies that reinforce women participation and foster a more diverse, egalitarian, and innovative technological landscape. As emerging technologies continue to shape the fu-ture, it is crucial that women are actively involved in shaping that future and in ensuring that digital technologies contribute to a more just and non-discriminatory world for all.

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. Bangun, R., & Fikri, S. (2024). The Role of Generative AI in Shaping Human Rights and Gender Equity: A Critical Analysis. Journal of Indonesian Legal Studies, 9(2), 799–834. https://doi.org/10.15294/jils.v9i2.13617
2. Battista, D., & Molano, J. C. (2023). How AI Bots Have Reinforced Gender Bias in Hate Speech. Ex Aequo - Revista Da Associação Portuguesa de Estudos Sobre as Mulheres, 48. https://doi.org/10.22355/exaequo.2023.48.05
3. Breda, T., Jouini, E., Napp, C., & Thebault, G. (2008). Gender stereotypes can explain the gender-equality paradox. PNAS, 117(49), 31063–31069. https://doi.org/10.1073/pnas.2008704117/-/DCSupplemental
4. European Commission. (2025). She figures 2024 : gender in research and innovation : statis-tics and indicators. Publications Office of the EU. https://op.europa.eu/en/publication-detail/-/publication/7646222f-e82b-11ef-b5e9-01aa75ed71a1/language-en
5. Eurostat. (2024). ICT specialists in employment. Eurostat. https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/47162.pdf
6. Guevara-Gómez, A., De Zárate-Alcarazo, L. O., & Criado, J. I. (2021). Feminist per-spec-tives to artificial intelligence: Comparing the policy frames of the European Union and Spain. Information Polity, 26(2), 173–192. https://doi.org/10.3233/IP-200299

7. ILO. (2019). The business case for change. International Labour Organization (No. 978-92-2-133168–1). International Labour Office. https://www.ilo.org/publications/women-business-and-management-business-case-change
8. Joyce, K. A., Darfler, K., George, D., Ludwig, J., & Unsworth, K. (2018). Engaging STEM Ethics Education. Engaging Science, Technology, and Society, 4, 1–7. https://doi.org/10.17351/ests2018.221
9. Klein, L., & D'Ignazio, C. (2024). Data Feminism for AI. The 2024 ACM Conference on Fairness, Accountability, and Transparency, 100–112. https://doi.org/10.1145/3630106.3658543
10. Lagesen, V. A., Pettersen, I., & Berg, L. (2022). Inclusion of women to ICT engineering–lessons learned. European Journal of Engineering Education, 47(3), 467–482. https://doi.org/10.1080/03043797.2021.1983774
11. Leslie, D., Katell, M., Aitken, M., Singh, J., Briggs, M., Powell, R., Rincón, C., Chengeta, T., Birhane, A., Perini, A., Jayadeva, S., & Mazumder, A. (2024). Advancing Data Justice Research and Practice: An Integrated Literature review. Zenodo. https://doi.org/10.5281/zenodo.6408304
12. McKinnon, M. (2020). The absence of evidence of the effectiveness of Australian gender equity in STEM initiatives. Australian Journal of Social Issues, 57(1), 202–214. https://doi.org/10.1002/ajs4.142
13. Scheele, A. (2005). The future of work – what kind of work? Impacts of gender on the defi-nition of work and research methodology. Transfer European Review of Labour and Re-search, 11(1), 014–025. https://doi.org/10.1177/102425890501100104
14. UNESCO. (2024a). Challenging Systematic Prejudices An Investigation into Gender Bias in Large Language Models. https://unesdoc.unesco.org/ark:/48223/pf0000388971
15. UNESCO. (2024b). Consultation paper on AI regulation: emerging approaches across the world. https://unesdoc.unesco.org/ark:/48223/pf0000390979

BRIDGING ETHICS AND GENDER EQUALITY IN AI EDUCATION: INSIGHTS FROM PORTUGAL

Camila Marques[1] [0009-0005-7873-1836], Rosa Monteiro[2] [0000-0002-2429-5590], Nuno Lourenço[3] [0000-0002-2154-0642]

Keywords: Higher Education, Gender, Ethics, Artificial Intelligence

Extended Abstract

Artificial Intelligence (AI) has the capacity to revolutionize numerous sectors of society, potentially impacting all aspects of life. However, research has shown that AI systems are not inherently neutral. They reflect and often reinforce existing social biases, which raise significant risks, particularly the perpetuation and exacerbation of existing inequalities, especially those related to gender and other underrepresented groups. As AI technologies continue to proliferate globally, it is imperative to ensure that their development, implementation, and regulation adhere to ethical and responsible practices that promote fairness, inclusion, and accountability.

Integrating ethics and gender perspectives into AI education represents a key strategy for fostering a more responsible technological landscape. By embedding these crucial considerations into the curriculum, we can ensure that future AI developers have the knowledge and skills to understand the importance of equality and ethical considerations in technology. This educational approach enables students, professors, and practitioners to critically analyze the societal impacts of AI systems, empowering them to make informed decisions that contribute to the development and deployment of socially responsible AI technologies.

The need to integrate ethics and gender perspectives into AI education aligns with the overarching objective of the present study, which is part of the project titled "Social Gender Relations in Change: Higher Education and Research Policies and Teaching/Learning Program Contents in Higher Education in Portugal". This project is being conducted by the Portuguese Women's Studies Association (APEM) and receives support from the Commission for Citizenship and Gender Equality (CIG). In addition, this study is part of a master's research project in Sociology at the Faculty of Economics of the University of Coimbra (FEUC), conducted by the student Camila Marques, under the scientific coordination and guidance of Professor Rosa Monteiro.

1 Faculty of Economics of the University of Coimbra, Av. Dias da Silva, 165 3004-512 Coimbra, Portugal, uc2020227742@student.uc.pt

2 Faculty of Economics of the University of Coimbra, Av. Dias da Silva, 165 3004-512 Coimbra, Portugal, rosa.monteiro@fe.uc.pt

3 Centre for Informatics and Systems of the University of Coimbra, R. Miguel Bombarda, 3030-790 Coimbra, naml@dei.uc.pt

This pioneering investigation aims to reflect on the integration of ethics and gender equality in the teaching of Artificial Intelligence (AI) in Portugal. In particular, we aim to identify if and how these questions are addressed in the courses of AI, considering where in the program they are included, whether they are addressed in an ad hoc or cross-cutting manner, which teaching methods are used, and what specific contents are covered. In addition, we aim to identify the initiatives being held by higher education institutions, professors, and other organizations in Portugal to combat gender inequalities in AI and contribute to the development of ethical and responsible AI systems.

The potential for AI to perpetuate and exacerbate inequalities, particularly gender disparities, is a pressing issue that has been extensively documented. Research shows that biased datasets, a lack of diversity in development teams, and uncritical approaches to algorithmic design can reinforce existing inequalities.

Addressing the risks associated with AI in education is crucial to promoting a more ethical development of this technology. Research has identified several good practices to help mitigate these challenges, including the integration of case studies that illustrate real-world challenges and the teaching of computational techniques, such as bias detection methodologies [1, 2, 5, 10, 11, 13]. Additionally, there has been a widespread discussion about the importance of using an interdisciplinary approach and integrating sociotechnical perspectives, drawing concepts from ethics and social sciences to critically analyze the interaction between technology and society [3, 6–9]. Our study in Portugal will reference comparative studies that report experiences of pedagogical approaches to ethics and gender equality in AI education [4, 12]. These studies emphasize interdisciplinary methodologies, scenario-based learning, and participatory approaches to foster critical thinking on bias, fairness, and inclusivity in AI development, and also how students perceive the importance of these topics in their education [12].

The present study employs a qualitative multi-phase methodology to analyze the integration of ethics and gender equality in AI education in Portugal. It combines a desk survey, an online questionnaire, and interviews.

The goal of the desk survey is to analyze the presence of ethics and gender equality perspectives in AI-related university degrees and courses offered by higher education institutions in Portugal. The analysis is based on information for the 2024/2025 academic year (or the most recent available data) gathered from institutional websites.

To carry out the analyses, a database was previously created to refine the selection of relevant university degrees. Creating this database involved several filtrations using the National Classification of Education and Training Areas (CNAEF).

Firstly, all AI-related degrees were mapped, focusing on fields like Computer Science, Informatics, Biotechnology, Bioinformatics, and Biomedical Engineering. Through subsequent filtrations, the list of university degrees was narrowed down to include only those related to Informatics, Computer Science, Life Sciences, and Electronics and Automation. A final filtration focused on identifying AI and Data Science programs, which led to the selection of 37 relevant courses across the three academic cycles (undergraduate, master's, and doctoral levels) for an in-depth analysis of their curriculums.

We broadened the universe in a new search with a different level of systematicity. We included all the second and third cycle university degrees, which mainly focus on computer science, biotechnology, bioinformatics, and biomedical engineering. We aim

to check whether these generic university degrees, but where AI issues are taught, also have courses on ethics and responsibility, and analyze their content.

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 of AI and Data Science university degrees and professors of AI degrees, including those teaching AI and ethics or responsibility. The questionnaire sought to assess faculty perspectives on the importance of including gender issues in AI education and how these topics could contribute to the development of future AI professionals capable of designing ethical and responsible AI systems.

In addition to the documentary analysis and survey, qualitative interviews are being conducted with key experts in AI and ethics, including professors and researchers. These interviews will provide 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.

Ultimately, this research highlights the need for a comprehensive and forward-thinking approach to AI education—one that extends beyond technical skills to critically engage with the social, ethical, and gender-related implications of AI. As AI increasingly shapes the fabric of society, its education must prepare not only skilled engineers but also responsible and conscientious developers who are equipped to navigate the broader societal impact of the technologies they create.

By analysing the integration of ethics and gender equality in AI education in Portugal, this study aims to showcase effective practices that could be adopted by higher education institutions, while identifying challenges that must be addressed. Furthermore, it will contribute to the ongoing dialogue about ethical AI development, offering practical recommendations for incorporating a gender perspective into AI programs in higher education.

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

References

1. 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.
2. CE: The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. Publications Office, LU (2020). https://doi.org/10.2759/791819.
3. 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.

4. 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.
5. 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.
6. Hoffman, S.G. et al.: Five Big Ideas About AI. Contexts. 21, 3, 8–15 (2022). https://doi.org/10.1177/15365042221114975.
7. Joyce, K. et al.: Engaging STEM Ethics Education. Engaging STS. 4, 1–7 (2018). https://doi.org/10.17351/ests2018.221.
8. Joyce, K. et al.: Toward a Sociology of Artificial Intelligence: A Call for Research on Inequalities and Structural Change. Socius: Sociological Research for a Dynamic World. 7, 2378023121999581 (2021). https://doi.org/10.1177/2378023121999581.
9. 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.
10. UNESCO: Ethical impact assessment. A tool of the Recommendation on the Ethics of Artificial Intelligence. UNESCO (2023). https://doi.org/10.54678/YTSA7796.
11. UNESCO: Metodologia de Avaliação de Prontidão: Um recurso da Recomendação sobre a Ética da Inteligência Artificial. (2023).
12. Villani, C. et al.: For a meaningful artificial intelligence: towards a french and european strategy. Conseil national du numérique (2018).
13. Zawacki-Richter, O. et al.: Systematic review of research on artificial intelligence applications in higher education – where are the educators? Int J Educ Technol High Educ. 16, 1, 39 (2019). https://doi.org/10.1186/s41239-019-0171-0.

GENDER AND EMERGING DIGITAL TECHNOLOGIES IN EDUCATION

Davorka Radaković[1] [0000-0001-8480-3211], William Steingartner[2] [0000-0002-2852-9403]

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

Extended abstract

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 [4],[5]; in programming [8]; creating content, answering questions, explaining concepts, and giving suggestions [6],[10].

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 [3]. 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.

Our literature review shows that researchers have extensively examined the potential of ChatGPT in teaching and learning in various subjects.

Qualitative research [7] provides an overview of the effectiveness of using ChatGPT as a virtual assistant. It helps improve student productivity in higher education

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

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

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 [3] 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 [2] 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 [11] 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 [12] 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 [5], 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.

Further, the study [9] found that students viewed ChatGPT as an innovative AI toolset applicable across classroom contexts, including programming and computer languages, promoting sustainable AI adoption in education.

Our study applies a cross-sectional analysis to examine students' attitudes toward AI in education. The research focuses on two key aspects of AI use, specifically through ChatGPT tool. As the first step, we examine whether and how students recognize AI as useful in their learning/studying process. Second, we explore their expectations, concerns, and potential barriers to adopting AIpowered tools in education. Our data collection methods include surveys and structured interviews, allowing us to gain insights into students' experiences, perceived benefits, and challenges.

Our findings report diverse perspectives on AI integration in education. While many students confirm and acknowledge the benefits of AI, including assistance in programming, language learning, and content creation, concerns persist regarding misinformation, cognitive overload, and mostly ethical considerations. The study highlights the importance of equipping educators and students with the necessary skills to critically assess AI-generated content. Furthermore, institutional support and policy frameworks are essential for the responsible and effective adoption of AI in teaching and learning. These results contribute to ongoing discussions on AI-driven education, emphasizing the

need for further research into best practices, ethical safeguards, and the long-term impact of AI-assisted learning.

We identify some challenges that higher education institutions and students may face, and we consider potential research directions.

Collaborating on the development and implementation of AI-driven digital solutions helps students, teachers, and university management better understand the challenges and opportunities of digital transformation. This collective experience deepens their commitment to the project's goals and fosters a culture of continuous improvement within their institutions.

This research enhances our understanding of LLMs' potential, limitations, and their integration into e-assessment systems, pedagogical scenarios, and student instruction using GPT-based applications.

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/, last accessed 8 February 2025
2. Csaba Norbert Nagy: The impact of ChatGPT on student performance in highereducation, vol. 3737 (2024)
3. 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. Journal of AI 7(1), 52–62 (Dec 2023)
4. 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 (jan 2025). https://doi.org/10.36732/riep.v7i1.421, https://ojs.novapaideia.org/index.php/RIEP/article/view/421
5. Çildir, M.: Collaborating with ChatGPT: An exploratory study of german languagelearning. Alman Dili ve Kültürü Araştırmaları Dergisi 5(2), 63–76 (Dec 2023)
6. 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) (Apr 2023)
7. Fauzi, F., Tuhuteru, L., Sampe, F., Ausat, A.M.A., Hatta, H.R.: Analysing therole of ChatGPT in improving student productivity in higher education. joe 5(4), 14886–14891 (Apr 2023)
4 D. Radaković et al.
8. Padilla, J.R.C., Montefalcon, M.D.L., Hernandez, A.A.: Language ai in programming: A case study of chatgpt in higher eduation 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
9. 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 (Feb 2024)
10. Thorp, H.H. Chatgpt is fun, but not an author. Science 379(6630), 313–13(2023). https://doi.org/10.1126/science.adg7879, https://www.science.org/doi/abs/10.1126/science.adg7879

11. Wu, Y.: Integrating generative AI in education: How ChatGPT brings challengesfor future learning and teaching. Journal of Advanced Research in Education 2(4), 6–10 (Jul 2023), https://doi.org/10.56397/JARE.2023.07.02
12. Zheng, Y.: Chatgpt for teaching and learning: An experience from data scienceeducation. In: Proceedings of the 24th Annual Conference on Information Technology Education. p. 66–72. SIGITE '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3585059.3611431

JUST HALLUCINATIONS? THE PROBLEM OF AI LITERACY WITH A NEW DIGITAL DIVIDE

Eleonora Bassi[1] [0000-0002-1619-3543], Ugo Pagallo[2] [0000-0001-7981-8849]

Keywords: Artificial Intelligence, Digital Divide, Literacy, Hallucination.

Extended Abstract

The Cambridge dictionary chose "hallucination" as the word of the year in 2023. The traditional definition of "hallucinate" was complemented with the notion of AI systems — in particular, Large Language Models, or in the jargon of EU law, General Purpose AI system, or model — which produce false information about the states of the world. The popularity of the neologism has gone hand in hand with a growing attention to the issues of AI literacy among scholars [1, 2, 3]; and institutions [4, 5, 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 the threat, from a new wave of Chat GPT-generated students' theses to attorneys providing memorandum of law on the basis of non-existent cases, quotes, and holdings, as occurred before the Southern District of New York in May 2023 (case Mata v. Avianca, no. 22-cv-01461). Against this framework, the claim of this paper is that a new digital divide, even harsher than the first wave in the early 2000s [7, 8], may follow as a result of today's AI illiteracy, in particular, of LLMs- or GPAI-related analphabetism.

The claim is twofold. On the one hand, 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. On the other hand, the uniqueness of AI — and moreover, of LLMs or GPAI technologies [9] — may bring forth new problems of access, usage, and outcomes, such as data inequalities, algorithmic awareness, misinformation hazards which affect trust, or perpetuation of biases and stereotypes triggering discrimination and hence, representational and allocational

1 Nexa Center, Polytechnic of Turin, Italy
2 Department of Jurisprudence, University of Turin, Italy

harms [10]. Correspondingly, the attention should be drawn not only to the cumulative effect of multiple waves of digital divide, but also, to that which is unique to the digital divide brought forth by new threats of AI illiteracy. For example, AI-generated fake news are often more credible to human raters than human-written disinformation, up to the point that texts, images, audios, and videos generated by LLMs can be impossible to distinguish from human creation [11]. This scenario clearly affects the normative and epistemic features of both AI literacy and social inclusion because inconclusive evidence can lead to unjustified actions as much as unfair outcomes can end up with discrimination or transformative effects that pose challenges to human autonomy [12, 13].

At the institutional level, both in the US and the EU for example, lawmakers have 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, so that individuals should understand how AI decisions may impact them in order "to take an active part in the economy, society, and in democratic processes" (Recital 56 of the AI Act). AI 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. Re-garding the means of regulation, the attention of lawmakers regards the challenges of AI literacy as a condition for the development of all those sectors — from the digital-isation of the public administration to the overall competitiveness of companies — which need new skills, abilities, and expertise [14].

However, in addition to the pressing problems of the present, focus should 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 [15, 16]. 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 looming on the horizon with new representational harms, information hazards, misinformation threats, or people overvaluing the capabilities of technology. The problem is not only the hallucinations of AI systems, but also, individuals who are not even capable of understanding they are facing just 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. Cambridge, MA: MIT Press (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, 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, available at https://en.unesco.org/futuresliteracy/about (2021)
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 information review, 25(5): 311-320 (2001)
8. Van Dijk, J. A., Digital divide research, achievements and shortcomings. Poetics, 34(4-5): 221-235 (2006)
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 Con-ference on Fairness, Accountability, and Transparency, 214–229 (2021)
11. Nightingale, S. J., & Farid, H., AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the National Academy of Sciences, 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 La politica dei dati. Il governo delle nuove tecnologie tra diritto, economia e società, edited by M. Durante and U. Pagallo, Milan, Mimesis (2022)
15. Yussen, S. R., The role of metacognition in contemporary theories of cognitive de-velopment. In Contemporary Research in Cognition and Metacognition, edited by D. For-rest-Pressley & G. Waller, Academic Press (1985)
16. Miller, R., Futures literacy: transforming the future. In Transforming the future. Anticipation in the 21st century, London/New York, Routledge Taylor & Francis Group, 1-12 (2018)

RESEARCH-BASED THEATER AS AN ENGAGING AND EFFECTIVE METHOD FOR PROMOTING AI LITERACY? – INSIGHTS FROM AUDIENCE RESPONSES

FRANZISKA POSZLER[1] [0000-0002-8135-6062]

Keywords: AI ethics, large language models, moral guidance, creative science communication, research-based theater

Extended Abstract

AI's societal implications & the need for AI literacy

Generative artificial intelligence (AI), especially chatbots powered by large language models (LLMs), enables individuals to increasingly rely on automation to support their decisions by answering questions, offering information or even providing concrete moral guidance [1], for instance, in the contexts of healthcare, criminal justice or personal matters [2]. Such systems, used to support decisions – especially those that ultimately affect core human values such as 'safety' or 'fairness' – can have far-reaching societal implications, including issues like responsibility gaps or moral deskilling [3]. Given the societal significance of these developed AI applications, it is particularly important to inform the public about the results of pertinent scientific research, a point emphasized in the open science focus areas [4] and recognized by regulatory initiatives. For example, the Pact for Research and Innovation IV (2021–2030) emphasizes science communication as a key objective for research institutions at the national level in Germany [5], while the EU AI Act stresses the importance of taking measures to ensure AI literacy on a global level [6]. Universities and research institutions, as key knowledge contributors, hold a unique responsibility in this endeavor [5].

Limits of traditional science communication & innovative methods of knowledge transfer

Science communication, especially in technical domains, faces various challenges, such as the often difficult-to-understand technical language in academic publications, which can reduce the public's interest in scientific texts [7]. To make scientific debates accessible beyond the academic community, innovative methods and non-traditional venues are needed through which civil society can stay informed about up-to-date research and engage with scientists [8; 9]. In this endeavor, arts – an important reference for social

1 INSTITUTE FOR ETHICS IN ARTIFICIAL INTELLIGENCE, TECHNICAL UNIVERSITY OF MUNICH (TUM), ARCISSTRASSE 21, 80333 MÜNCHEN, GERMANY FRANZISKA.POSZLER@TUM.DE

knowledge and inclusion – can become a key enabler to ensure human-centric, participatory discussions around the design of pertinent technological innovations [10]. More specifically, arts-based methodologies such as research-based theater, ethnodrama or the playwright-approach "combine research and theater to create novel opportunities for inquiry and knowledge translation" [11; p.1). Especially in the context of AI, the use of theater is valuable to facilitate enlightenment about and the human-centricity of emerging technologies [12]. First pertinent efforts are made in the field of human-computer interaction, where researchers have started to utilize immersive theater performances to test prototypes, rapidly learn about future technology and design it accordingly [13].

Although initiatives like these are fundamentally aimed at expanding societal knowledge in the domain of AI, the explicit evaluation of the effectiveness of knowledge transfer is often neglected. Without evaluating the actual knowledge gained by the audience, the contribution of such arts-based approaches to fostering AI literacy remains merely an assumption. Therefore, this study aims to assess the effectiveness of an exemplary research-based theater initiative called 'MoralPLai' by conducting accompanying research through post-performance audience surveys and interviews to ultimately address the following key question:

- How effectively can research-based theater be used as a tool to promote a participatory approach in AI ethics research and enhance AI literacy among the audience?

The MoralPLai projecto & its accompanying research

The 'MoralPLai' project[2] implements a creative approach to conducting, educating, and communicating AI ethics research through the lens of the arts (i.e., research-based theater). The core idea revolves around conducting qualitative interviews and user studies on the impact of large language model(LLM)-based chatbots on human ethical decision-making. It focuses specifically on exploring the potential opportunities and risks of employing these systems as aids for ethical decision-making, along with their broader societal impacts and recommended system requirements. Generated scientific findings will be translated into a theater play, which will be performed in the theater hall in Germany. This performance seeks to effectively educate civil society on up-to-date research in an engaging manner and facilitate joint discussions (e.g., on necessary and preferred system requirements or restrictions). The insights from these discussions, in turn, are intended to inform the scientific community, thereby facilitating a human-centered development and use of LLMs as moral dialogue partners or advisors.

To evaluate the effectiveness of the MoralPLai performance, accompanying research will be conducted using mixed methods. Namely, audience responses will be captured through a post-performance survey including questions scored on a Likert scale with a space for open-ended comments as well as through semi-structured interviews via Zoom (adopted from [14]). The questions posed to the audience will center on how their understanding, learning, and awareness of the responsible use of LLM-based chatbots

2 Link to MoralPLai webpage: https://www.ieai.sot.tum.de/research/moralplai/

as moral advisors have changed after watching the play. Additionally, the questions will explore their views on the effectiveness of the play as a tool for promoting AI literacy, their preferred system requirements for developing LLM-based chatbots as well as demographic details such as age and occupation. The collected data will be analyzed using a mix of quantitative statistical analyses and qualitative content analysis [15].

Conclusion

To summarize, through post-performance audience surveys and interviews, this study aims to shed light on how much the MoralPLai performance influenced the audience's opinions, knowledge, and awareness regarding the societal implications of pertinent AI systems as well as on their preferred system requirements and the relevant values underlying the development of LLM-based chatbots. Generated findings can inform technology companies about appropriate and socially preferred system requirements to consider in order to ensure the development of trustworthy systems. For scholars and theater practitioners, the findings generated and the evaluation measures used in this study can serve as a foundation and proof of concept to legitimize and encourage the adoption of similar arts-based research projects and creative methods for science communication in the future. Overall, the MoralPLai project and its accompanying research may constitute the beginning of transforming scientific communication and inquiry in the field of AI ethics, with the humanities and cultural sciences serving as critical methodologies to bridge the gap between academia and society and promote AI literacy.

Acknowledgments: The MoralPLai project and this study was funded by the Notre Dame-IBM Tech Ethics Lab, the Friedrich-Schiedel-Fellowship Program at the TUM Think Tank and the TUM School of Social Sciences and Technology, the TUM Global Incentive Fund, the Tuition Substitution Funds, the Klaus Tschira Foundation gGmbH, TUM Center for Culture and Arts (CCA) as well as the TUM Global Visiting Professor Program.

References

1. Aharoni, E., Fernandes, S., Brady, D. J., Alexander, C., Criner, M., Queen, K., ... & Crespo, V.: Attributions toward artificial agents in a modified Moral Turing Test. Scientific Reports 14(1), 8458 (2024)
2. Awad, E., & Levine, S.: Why we should crowdsource AI ethics (and how to do so responsibly). Behavioral Scientist, (2020)
3. Poszler, F.: Integrating ethical principles into AI systems: Practical implementation and societal implications (Doctoral dissertation). Technical University of Munich, Munich (2024)
4. Ramachandran, R., Bugbee, K., & Murphy, K.: From open data to open science. Earth and Space Science 8(5), 1-17 (2021)
5. Wissenschaftsrat: Wissenschaftskommunikation: Positionspapier, https://www.wissenschaftsrat.de/download/2021/9367-21.pdf?__blob=publicationFile&v=10, last accessed 2025/03/29 (2021)
6. European Commission: REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 June 2024 laying down harmonised rules on artificial intelligence, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689, last accessed 2025/03/29 (2024)
7. Dietrich, A., Liepelt, R., & Sperl, L.: Die Wissenschaftsautor* innen Ihres Vertrauens-Über die Hürden von Wissenschaftskommunikation. The Inquisitive Mind, (2024)

8. Wissenschaftlicher Beirat der Bundesregierung Globale Umweltveränderungen (WBGU): Unsere gemeinsame digitale Zukunft – Empfehlungen, https://www.wbgu.de/fileadmin/user_upload/wbgu/publikationen/hauptgutachten/hg2019/pdf/WBGU_HGD2019_Empfehlungen.pdf, last accessed 2025/03/29 (2019)
9. Weber, C., Allen, S., & Nadkarni, N.: Scaling training to support scientists to engage with the public in non-traditional venues. Journal of Science Communication 20(04), 1-15 (2021)
10. Guryanova, A. V., & Smotrova, I. V.: Transformation of worldview orientations in the digital era: humanism vs. anti-, post-and trans-humanism. In: International Scientific Conference "Digital Transformation of the Economy: Challenges, Trends, New Opportunities", pp. 47-53. Springer, Cham (2019)
11. Nichols, J., Cox, S. M., & Guillemin, M.: Road Trips and Guideposts: Identifying and Navigating Ethical Tensions in Research-Based Theater. Qualitative Inquiry 29(2), 257-266 (2022)
12. Fukuyama, M.: Society 5.0: Aiming for a new human-centered society. Japan Spotlight 27(5), 47-50 (2018)
13. Luria, M., Oden Choi, J., Karp, R. G., Zimmerman, J., & Forlizzi, J.: Robotic Futures: Learning about Personally-Owned Agents through Performance. In: Proceedings of the 2020 ACM Designing Interactive Systems Conference, pp. 165-177. Eindhoven (2020)
14. Belliveau, G., & Nichols, J.: Audience responses to Contact! Unload: A Canadian research-based play about returning military veterans. Cogent Arts & Humanities 4(1), 1-12 (2017)
15. Gioia, D. A., Corley, K. G., & Hamilton, A. L.: Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. Organizational research methods 16(1), 15-31 (2013)

ARTIFICIAL INTELLIGENCE IN SCIENTIFIC RESEARCH: THE MISSING LINK OF LITERACY FOR SECURITY CONCERNS

LUDOVICA PASERI[1] [0000-0002-5818-7969]

Keywords: Artificial Intelligence, AI Act, AI literacy, Scientific research, Research security

In May 2024 the Council of the European Union issued the Council Recommendation on enhancing research security stating that "with growing international tensions and the increasing geopolitical relevance of research and innovation, the Union's researchers and academics are increasingly exposed to risks to research security when cooperating internationally" (Council of the EU, 2024, 2). The Council of the EU deems it necessary to act at the EU level in order to "protect the integrity of the ERA [European Research Area], while respecting the competences of Member States for going further, for example by developing regulatory frameworks" (Council of the EU, 2024, 9). In addressing these concerns, the Recommendation emphasises the importance of balancing the openness of international collaboration (Paseri, 2024a) with robust measures to safeguard the security of research. Although the EU Council proposes a broad definition of "research security" (Council of the EU, 2024, 18), this aspect requires further investigation to understand the relationship between the notion with that of "data security", and potential intersections with the concept of cybersecurity (Shih, 2025; Shankar and Drake 2025). The Recommendation supports the establishment of common standards and guidelines for all Member States to address risks such as intellectual property theft, misuse of research results and interference by foreign actors. The Council aims to promote a resilient ERA that can thrive in an increasingly complex and competitive global landscape, avoiding "risk of undesirable transfer of critical knowledge and technology […] affecting the security of the Union and its Member States" (Council of the EU, 2024, 4).

The Council Recommendation on enhancing research security identifies artificial intelligence (AI, hereinafter) as a matter of priority with other three critical technology areas (i.e., semiconductors, quantum, and biotechnologies). This is the case both when considering an AI system as a tool and as a result of a research project. In the former case, an AI system is used by researchers in order to implement a specific research project, becoming instrumental in obtaining the results (Paseri, 2024, 59). In the second case, artificial intelligence is the area of investigation and the AI system is the result of a research project (Paseri, 2024, 60). Both the use and development of AI systems in the context of scientific research represent sensitive scenarios due to the high stakes involved. Thus far, several studies have been conducted on the impact of AI and generative AI (gen AI, hereinafter) on openness and collaboration in science (Beck et al., 2022; Wang

1 LAW DEPARTMENT, UNIVERSITY OF TURIN, ITALY, LUDOVICA.PASERI@UNITO.IT

and Barabási, 2021). The positions on the issue tend to be polarized – in Umberto Eco's words – between apocalyptic and integrated intellectuals (Eco, 1964). On the one hand, over-enthusiastic scholars have gone so far as to claim that "by harnessing the power of AI we can propel humanity toward a future where groundbreaking achievements in science, even achievements worthy of a Nobel Prize, can be fully automated", believing "that this is achievable by the year 2050" (King et al., 2024, 716). On the other hand, more and more scientific journals have updated their ethical guidelines identifying the use of undeclared AI or gen AI as forms of scientific misconduct (Kumar et al., 2024).

In order to handle this sensitive matter and to guarantee research security, the Council first recalls the framework of the EU Security Union Strategy, which proposes a three-pronged approach, on "promotion of the Union's economic base and competitiveness; protection against risks; and partnership with the broadest possible range of countries to address shared concerns and interests" (EU Security Union Strategy, 2023). Further, the Council proposes a series of recommendations addressed to research entities; to the Member States called upon to act at the regulatory level; and to the European Commission in its coordinating role in the field of scientific research policies (Council of the EU, 2024, 13-16, 19, 23).

Against this backdrop, what is lacking in the debate on research security and use of AI in science is literacy, intended as the ability "in both the human and technological dimensions of AI, understanding how it works in broad terms, as well as the specific impact of Gen AI" (UNESCO, 2023, 24). In addition to the shortcomings of the Artificial Intelligence Act (AI Act, 2024) (Pagallo, 2024, 64-65), it is worth noting that the European lawmaker pays attention to the issue of AI literacy. As stated in Article 4 of the AI Act, "Providers and deployers of AI systems shall take measures to ensure, to their best extent, a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems on their behalf, taking into account their technical knowledge, experience, education and training and the context the AI systems are to be used in, and considering the persons or groups of persons on whom the AI systems are to be used". This provision is particularly relevant for the use of AI in scientific research, where the complexity of AI systems demands a high level of expertise among researchers and operators. The problems are both epistemic and normative (Durante and Paseri, 2005). Ensuring AI literacy within research teams helps to enhance the reliability and validity of scientific findings, as well as to address ethical considerations and biases related to AI models. By aligning with the requirements of the AI Act, research institutions can foster responsible and informed use of AI technologies, ultimately advancing innovation while safeguarding integrity and accountability in scientific inquiry. Nevertheless, it is necessary to investigate what is meant by AI literacy applied to the field of scientific research and how it can be implemented or realised. There are, for instance, universities that have published guidelines on the ethical use of AI (e.g., University of Milan, 2025), and on the other hand research organisations that have banned its use.

The paper points out the lack of consideration about the AI literacy in the context of research security and intends to investigate the potential inequities in research resulting from the AI illiteracy (e.g., loss of funding opportunities; methodological disadvantages; bias in AI-driven tools; institutional disparities, etc.). Strengthening research security aspects and adopting a polarised approach about the role of AI in science,

among other things, risks acquiring inequities and undermining inclusiveness in research, a key factor in EU policies on science and a priority frequently evoked by scholars (Rafols et al. 2024). As Stefano Rodotà emphasised with regard to the risks associated with the digital divide, selectively benefiting from technological innovation "leads to a «human divide»" (Rodotà, 2012, 198), which poses an even greater risk when applied in the context of scientific inquiry.

References

1. Beck, S., Poetz, M., and Sauermann, H. (2022) "How will Artificial Intelligence (AI) influence openness and collaboration in science?", Elephant in the Lab, https://research.cbs.dk/en/publications/how-will-artificial-intelligence-ai-influence-openness-and-collab.
2. Council of the European Union, Council Recommendation on enhancing research security, 2024/0012(NLE), Brussels, 14 May 2024.
3. Eco, U. (1964), Apocalittici e integrati: comunicazioni di massa e teorie della cultura di massa, Bompiani, Milano.
4. EU Commission, Joint communication to the European Parliament, the European Council and the Council on "European Economic Security Strategy", JOIN(2023) 20 final, ELI: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52023JC0020.
5. King, R.D., Scassa, T., Kramer, S., Kitano, H., "Stockholm declaration on AI ethics: why others should sign", Nature 626, 2024, pp. 716-716.
6. Kumar, R., et al. (2024) "Academic integrity and artificial intelligence: An overview", Second Handbook of Academic Integrity, pp. 1583-1596.
7. Pagallo, U. (2024) "Introduction to a Theory of Legal Monsters: From Greco Roman Teratology to the EU Artificial Intelligence Act", i-lex 17.1, pp. 53-73.
8. Paseri, L. (2024a), Scienza aperta. Politiche europee per un nuovo paradigma della ricerca, Mimesis edizioni, Milano-Udine.
9. Paseri, L. (2024b) "Science and Technology Studies, AI and the Research Sector: Questions of Identity", Elliott, A. (ed.), The De Gruyter Handbook of Artificial Intelligence, Identity and Technology Studies, De Gruyter, Berlin, pp. 55-72.
10. Paseri, L., Durante, M. (2025) "Examining epistemological challenges of large language models in law", Cambridge Forum on AI: Law and Governance, pp. 1-13.
11. Rafols, I., et al. (2024), "The multiversatory: fostering diversity and inclusion in research information by means of a multiple-perspective observatory", Conference on Advancing Social Justice Through Curriculum Realignment, pp. 1-15, https://osf.io/preprints/socarxiv/dn2ax.
12. Rodotà, S., Il diritto di avere diritti, Editori Laterza, Bari, 2012.
13. Shankar, A., Drake, W. (2022), Effective Cybersecurity for Research, Center for Applied Cybersecurity Research Indiana University, pp. 1-25.
14. Shih, T. (2025) "Challenges to research security", SSRN, pp. 1-8.
15. UNESCO. (2023). Guidance for generative AI in education and research. https://unesdoc.unesco.org/ark:/48223/pf0000386693.
16. University of Milan (2025), 10 principles AI, https://www.unimi.it/it/ateneo/normative/linee-guida/decalogo-un-utilizzo-etico-legittimo-e-consapevole-di-strumenti-dintelligenza-artificiale-ai-tutte.
17. Wang, D., Barabási, A. L. (2021) The science of science, CUP, Cambridge.

A REALISTIC APPROACH TO THE OBVIOUSNESS CONCEPT WITHIN THE AI ACT: A MATTER OF AI LITERACY

Jacopo Ciani Sciolla[1] [0000-0003-2707-4694]

Keywords: Artificial Intelligence, Deep Fake, Transparency, Literacy, Obviousness.

Extended Abstract

The European Union's Artificial Intelligence Act (AI Act) is a landmark regulatory framework that seeks to establish comprehensive guidelines for the development, deployment, and use of artificial intelligence (AI) technologies within the EU. As AI systems become increasingly pervasive across various sectors, ensuring their ethical, safe, and transparent use is crucial. One of the foundational elements of achieving these goals is fostering AI literacy—the ability to understand, interpret, and critically evaluate AI systems and their impacts. The concept of AI literacy is embedded within the AI Act, not only as an essential skill for stakeholders interacting with AI but also as a vital tool for ensuring public trust and accountability.

This extended abstract explores the relevance of AI literacy in the context of a key provision of the AI Act, namely the obligation to disclose AI generated contents, with the purpose of minimizing the "new risks of misinformation and manipulation at scale, fraud, impersonation and consumer deception" and restore "the integrity and trust in the information ecosystem" (recital 133). Drawing on the experience of many AI ethics charters and guidelines, setting transparency, articulated as the duty to make an object or entity knowable, as a pillar principle for any AI deployment, Article 50 para. 1 establishes a duty of disclosure concerning "providers" placing on the EU market or putting into service AI systems or general-purpose AI models, as defined in Art. 3(3). A provider shall ensure that "AI systems intended to interact directly with natural persons are designed and developed in a way that the natural persons concerned are informed that they are interacting with an AI system".

This provision directly concerns a myriad of virtual influencers, avatars, chatbots, and other non-human digitally created characters populating social media, sharing content, and engaging in interactive communications with consumers. They may have many different forms to the point that some authors developed a taxonomy based on their similarity to human appearance, also known as anthropomorphism, and their placement on the reality-virtuality continuum, ranging from unimaginable to hyper-realistic characters that can be nearly impossible to distinguish from humans.

1 University of Turin, Department of law, Turin 10121, Italy, jacopo.cianisciolla@unito.it

The duty to disclose their artificial nature is not an absolute requirement. Under Article 50, natural persons should be notified that they are interacting with an AI system "unless this is obvious from the point of view of a natural person who is reasonably well-informed, observant and circumspect, taking into account the circumstances and the context of use."

In a nutshell, the AI Act acknowledges that in some cases, the nature of AI systems is so apparent that it doesn't require disclosure duties. Therefore the concept of "obviousness" plays a crucial role in determining the scope and applicability of compliance obligations. "obviousness" is used to delineate situations in which the risks associated with AI systems are sufficiently apparent or self-evident to stakeholders (e.g., users). This can influence which kind of transparency measures are considered necessary. For AI providers, the understanding of whether their system is "obviously AI" can significantly affect the resources and efforts required for compliance. If an AI system is deemed recognizable based on its obviousness, it must undergo less rigorous transparency measures.

To this extent, obviousness is a significant determinant of regulatory obligations. This approach is problematic because it leaves any evaluation to the providers and adds a layer of subjectivity and uncertainty. For example, AI systems having a long lasting history may be more easily recognizable as such than novel AI technologies operating in new or unforeseen contexts. The concept of "obviousness" is fraught with ambiguity and presents challenges in its application, given the dynamic and complex nature of AI technologies as well as the very uncertain level of AI literacy of the average consumer. Users—whether individuals, businesses, or public sector entities—may have a different level of AI literacy. This involves a combination of skills, knowledge, and awareness related to artificial intelligence systems, including technical understanding such as the knowledge of the fundamental principles behind AI technologies, such as machine learning, data processing, and algorithmic decision-making. At the same time, AI literacy involves the ability to recognize the ethical implications of AI, the capacity to assess the potential impacts of AI on individuals as well as the awareness of existing legal frameworks that govern the use of AI.

The legislator does not clarify the level of AI literacy of the average consumer, nor which aspects of AI literacy are relevant for the obviousness test.

This uncertainty may be particularly prejudicial also in consideration of Article 50(7) which encourages "the drawing up of codes of practice at Union level to facilitate the effective implementation of the obligations regarding the detection and labeling of artificially generated or manipulated content". Any ambiguity in the interpretation of the obviousness test may cause self-regulation authorities to set totally different standards at the national level. At the same time, platforms which are already taking matters into their own hands may develop different standards of implementation of the EU rules.

The European Committee of the Regions pointed out some of these concerns in its Opinion over the AI Act proposal, inviting to remove the exception on the grounds that "natural persons should always be duly informed whenever they encounter AI systems and this should not be subject to interpretation of a given situation. Their rights should be guaranteed at all times in interactions with AI systems". As the exception has been maintained, it is critical to start reflecting on it. While challenges exist in fostering

widespread literacy, particularly among non-expert stakeholders, the problem of the obviousness test has not been already tackled by scholars. It is therefore urgent to start developing a common approach to obviousness, to ensure that it accurately captures risks while promoting innovation and safeguarding fundamental rights. This is the main aim of the proposed paper. Since AI systems have the potential to impact fundamental rights, such as privacy, discrimination, and fairness, it is important to develop a clear, not overly simplistic, nor superficial understanding of what constitutes "obviousness" in the context of AI. At the same time, the adoption of a flexible, adaptive approach to defining "obviousness", might be better rather than relying on narrow or rigid conceptions.

All in all, the paper addresses the issue if the normative translation of the principle of transparency within the context of Article 50 is truly conducive to knowledge or risk not to be effectively put into practice. Then, the paper aims to give an example of how transparency obligations are framed in the AIAct as having a selective, rather than generalised application and shows how this may affect individual's rights. Third, the paper shall foster discussion on the criteria that should be relied on to assess fundamental rights violations so that the AIA's effective implementation is not compromised.

References

1. Voigt, O., Hullen, N.: The EU AI Act. Answer to Frequently Asked Questions Springer Berlin, Heidelberg (2024)
2. Nikolinakos, N. Th.: EU Policy and Legal Framework for Artificial Intelligence, Robotics and Relate Technologie – The AI Act Springer Cham (2023)
3. Romero Moreno, F.: Generative AI and deepfakes: a human rights approach to tackling harmful content. International Review of Law, Computers & Technology 38(3), 297-326 (2024)
4. Labuz, M.: A Teleological Interpretation of the Definition of Deep Fakes in the EU Artificial Intelligence Act – A Purpose -Based Approach to Potential Problems with the Word "Existing". Policy & Internet 17(1), 1–14 (2025)
5. Rijsbosch, B., van Dijck, G., Kollnig, K.: Adoption of Watermarking for Generative AI Systems in Practice and implications under the new EU AI Act arXiv:2503.18156 (2025)
6. Gils, T.: Article 50. In Pehlivan, C.N., Forgo, N. Valcke, P. (Eds.), The EU Artificial Intelligence (AI) Act: A Commentary. Kluwer Law Int., Alphen aan den Rijn (2024)
7. Meding, K., Sorge, C.: What constitutes a Deep Fake? The blurry line between legitimate processing and manipulation under the AI Act. In: CSLAW'25: Proceedings of the 2025 Symposium on Computer Science and Law, pp. 152–159. ACM, New York (2025)

NORMATIVE ISSUES OF AI LITERACY IN CONSTRUCTION PROJECT RISK MANAGEMENT

PLOTKA SAJJAD

Abstract

This research investigates AI literacy challenges in construction project risk management, finding that ethical, legal, and social factors outweigh technical barriers. Through analysis of nine industry professionals' insights, the study identifies four key domains: regulatory frameworks, knowledge management, socio-economic inequities, and ethical implementation with human oversight. Using theoretical frameworks including Institutional Theory and Value-Sensitive Design, the research reveals how regulations like the EU AI Act create compliance challenges, particularly for smaller firms. The study highlights how construction's reliance on tacit knowledge conflicts with AI implementation requirements, while resource disparities create unequal technology access across the industry. The research proposes four interventions: education framework development, practice evaluation, industry-wide initiatives, and democratic governance integration. It translates regulatory requirements into practical frameworks, providing guidance for responsible AI integration that maintains safety standards while addressing equity concerns. The findings emphasize that sustainable AI literacy requires collaborative approaches with worker participation and human oversight of AI decisions.

Introduction

The construction industry faces a critical challenge where traditional project risk management intersects with advancing machine learning technologies. While AI's potential for enhancing risk prediction and mitigation is increasingly recognized (Yao and Soto, 2022), successful integration extends beyond technical implementation. The industry's complex risk landscape creates unique challenges for AI adoption that go beyond conventional technological concerns (Chiu and Lim, 2021)

What initially was an investigation into ML integration across construction divisions revealed an essential challenge: the normative issues of AI literacy among professionals. Through analysis of transparency gaps, human oversight deficiencies, and training inadequacies under the EU AI Act and GDPR, significant compliance risks emerged (Musch et al., 2024). These investigations exposed biased risk models violating ethical protocols, highlighting AI literacy as a critical factor affecting effective ML integration in project risk management (Hannah et al., 2024).

Construction industry's traditional dependence on experimental knowledge slows technological changes (Gann, 2001). When combined with evolving regulatory requirements, this creates a complex normative landscape which industry must navigate. The EU AI Act's Article 4 requirement for "sufficient level of AI literacy" and the White House Executive Order's Section 10.2 addressing workforce development demand that construction organizations develop comprehensive literacy frameworks addressing both technical competencies and ethical responsibilities (Fernandes et al., 2024; Biden, 2023; Oladinrin and Ho, 2014)

The industry's fragmentation, from large companies to small local businesses, creates different competences for implementing AI literacy initiatives. This social stratification, combined with project-based work environments and diverse workforce demographics, presents unique challenges (Lloyd Walker et al., 2018) AI literacy in construction represents not merely a training challenge but a fundamental normative issue requiring systematic industry-wide transformation (Mittal et al., 2018)

Research Questions

This study addresses three interconnected research questions exploring the normative challenges of AI literacy implementation in construction project risk management:

1. How do current AI literacy levels and digital technology practices in the construction industry align with emerging regulatory requirements?
2. What are the key barriers and opportunities for implementing AI literacy requirements across different construction industry stakeholders?
3. What strategies and frameworks can be developed to bridge the gap between current industry practices and required ethical and legal standards in construction?

Significance:

This study was conducted because initial ML integration research unpredictably revealed normative challenges as the critical barrier along with technical limitations. When analysing EU AI Act, GDPR compliance and US White House Order highlighted AI literacy as a fundamental, overlooked issue requiring immediate attention (Butt, 2024) The study adds value by making regulatory requirements like EU AI Act Article 4 actionable within construction's unique context. Through analysis of nine industry professionals, it identifies four critical themes where normative challenges intersect with practical implementation, establishing under theoretical frameworks. The study delivers tangible recommendations on education, evaluation, industry initiatives, and governance that specifically address identified barriers while accommodating diverse organizational resources (Farahani, 2024), This bridges the gap between abstract regulatory requirements and implementable strategies, offering construction stakeholders evidence-based guidance for responsible AI integration that maintains safety standards while addressing equity concerns (Burr and Leslie, 2023).

Literature Review

AI and ML in Construction

Artificial Intelligence and Machine Learning applications in construction have transformed traditional project management approaches through predictive analytics, automated scheduling, and safety monitoring (Bilal et al., 2016; Datta et al., 2024). However, the construction industry is still considered one of the least digitally transformed sectors, with AI adoption rates significantly far behind other industries (Regona et al., 2022)

AI applications in construction project risk management (PRM) represents an emerging field categorized by fragmented employment and limited standardization (Secundo et al., 2024). While AI systems determine substantial potential for enhancing risk identification and mitigation processes, their deployment faces significant organizational barriers (Mohamed et al., 2025). The industry's traditional reliance on tacit knowledge creates fundamental tensions between established practices and emerging technological capabilities (Abbeasnejad et al., 2021).

Current applications predominantly utilize supervised learning for pattern recognition in historical project data, with emerging research exploring reinforcement learning for dynamic risk scenarios (Waqar et al., 2024). However, effectiveness remains dependent upon data quality, where construction organizations consistently struggle due to inadequate information management practices (Barbosa et al., 2018; Barbosa et al., 2023). This technological and organizational disconnect underscores the critical importance of AI literacy development within construction contexts.

AI Literacy

AI literacy has emerged as a critical competency framework encompassing both technical understanding and ethical awareness of artificial intelligence systems (Pinski, 2024)). Long and Magerko (2020) define AI literacy as "a set of competencies that enables individuals to critically evaluate AI technologies, communicate and collaborate effectively with AI, and use AI as a tool online, at home, and in the workplace."

Current literature divides AI literacy into operational proficiency and socio-technical comprehension, incorporating understanding of algorithmic bias, ethical implications, and societal impacts (Tadimalla and Maher, 2024). Ng et al. (2021) proposed a layered framework that shows AI literacy requires developing technical skills, emotional awareness and mental understanding.

The European Commission's AI Literacy Framework (2024) provides institutional recognition, establishing baseline competencies aligned with digital literacy initiatives. Parallel developments in the United States, including the White House Executive Order on AI (2023), explicitly address workforce development needs for responsible AI deployment. These policy frameworks signal systematic AI literacy requirements across professional domains, particularly in sectors involving public safety and critical infrastructure.

AI Literacy in the Construction Industry

Construction-specific AI literacy research remains limited, though emerging studies highlight unique challenges within the industry's traditional culture (Cheng et al., 2024). The workforce's demographics such as aging management and diverse technical backgrounds could create distinct development challenges not observed in tech-centric industries. Additionally, project-based work complicates systematic training initiatives across frequently changing organizational contexts.

Supranaviciene et al. (2024) identified construction-specific AI literacy gaps across three domains: technical understanding of AI applications in risk assessment, regulatory compliance awareness, and ethical considerations regarding worker data collection. Current professional development programs inadequately address these unique requirements, particularly regarding AI integration with existing risk management protocols.

Recent case studies demonstrate successful AI literacy initiatives require substantial organizational restructuring and cultural change management (Maitz et al., 2020). These implementations reveal that effective programs must address both individual competencies and collective organizational capabilities for knowledge management. The challenge intensifies for SMEs, which constitute the majority of construction firms but possess limited resources for comprehensive literacy development (Rivera et al., 2024; Saka and Chan, 2020).

Regulatory Frameworks and Standards

The regulatory landscape governing AI implementation in construction operates within an evolving framework of international standards and national legislation. The EU AI Act (Regulation 2024/1689) represents the most comprehensive regulatory approach to date, establishing risk-based classification systems and mandatory literacy requirements for AI system deployers. Article 4 specifically mandates sufficient AI literacy for staff, creating direct obligations for construction organizations implementing AI systems (Fernandes et al., 2024)

The White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI (2023) provides parallel frameworks in the United States, with Section 10.2 lett. (d) vii specifically addressing AI literacy requirements for organizations developing and deploying AI systems. These regulations create complex compliance landscapes for international construction firms navigating divergent US and EU requirements (Biden, 2023)

Beyond AI-specific legislation, the General Data Protection Regulation (GDPR) creates additional compliance considerations for construction organizations implementing AI systems that process worker and project data. The intersection of GDPR requirements with AI Act provisions establishes comprehensive data protection frameworks affecting how construction firms develop and deploy AI literacy programs (Regulation, 2018)

Professional standards organizations contribute additional frameworks, including ISO/IEC 42001 for AI management systems, ISO/IEC 23894 for risk management principles, and IEEE 7000-2021 for addressing ethical design considerations. The

European Commission's DigComp 2.2 framework provides foundational digital competency guidelines that intersect with AI literacy requirements. Additionally, emerging frameworks like the Digital Services Act (DSA) and Digital Markets Act (DMA) create broader digital ecosystem regulations that influence AI implementation contexts within construction organizations (Chiarella, 2023; Fernandes, 2025; Morandín-Ahuerma, 2023; Van Audenhove et al., 2024)

However, construction-specific applications of these frameworks remain limited, creating implementation challenges where organizations must adapt general guidelines to industry-specific operational contexts. This regulatory complexity, combined with the rapid pace of AI technological development, creates what legal scholars' term "implementation gaps" requiring organizations to develop interpretive approaches for compliance within their specific operational environments.

Methodology

This study adopted a philosophical stance bridging empirical observations with normative analysis, acknowledging that understanding AI literacy challenges in construction requires both practical insights and ethical considerations (Halder and Batra, 2024). The research design integrated real-world experiences with regulatory frameworks, recognizing that effective AI literacy goes beyond technical competency to encompass ethical responsibilities and compliance requirements (Ho et al., 2019).

Research Design

The study employed a qualitative approach, designed to capture the complex interaction between practical implementation challenges and normative requirements in AI literacy. Drawing from both constructivist and pragmatic paradigms, this design acknowledged that AI literacy understanding is socially constructed while remaining grounded in practical industry needs. The selection of participants from IT, Academia, and Industry sectors (n=9) from a pool of 73 professionals was intentional, ensuring diverse perspectives while maintaining focus on practical implementation challenges.

Data Collection

Semi-structured interviews served as the primary data collection method, allowing for both structured inquiry and exploratory discussions. (Karatsareas 2022) The interview protocol was designed to uncover not only current practices but also ethical considerations and normative challenges in AI literacy. Participant selection criteria included:

- Minimum 10 years of construction industry experience
- Involvement in ML implementation
- Decision-making authority in their organizations

This careful selection ensured that participants could speak to both practical challenges and normative implications of AI literacy in construction project risk manage

ment. Interviews averaged 45 minutes, conducted virtually through secure platforms, and were recorded with participant consent, maintaining ethical research standards.

Data Analysis

The analysis followed a thematic analysis approach (Braun & Clarke, 2006) with initial coding identifying 453 points of interest, refined to 323 significant quotes. Four primary themes emerged from the data with relevant theories and regulatory connections:

Theme	Core Theory	Regulatory Connection
Regulatory and Governance Frameworks	Institutional Theory (DiMaggio & Powell, 1983), Reflexive Regulation Theory (Gunningham, 2012)	White House Executive Order on AI (2023), EU AI Act Article 4 (Regulation 2024/1689) Section 10.2
Knowledge Management and Standardization	SECI Knowledge Model (Nonaka & Takeuchi, 1995)	EU AI Act Article 4 "sufficient level of AI literacy", White House Order systematic AI development requirements
Socio-Economic Inequities	Resource Dependence Theory (Pfeffer & Salancik, 1978)	White House Order democratic AI access objectives, EU AI Act human rights considerations
Ethical Implementation and Human Oversight	Value-Sensitive Design Theory (Friedman et al., 2013)	GDPR data protection requirements, EU AI Act Articles 13-14 (transparency and human oversight), White House Order "trustworthy" AI development

Analysis proceeded through iterative cycles connecting participant quotes with theoretical concepts and regulatory requirements. This synthesis revealed four recommendation domains: Education Framework Development, Professional Practice Evaluation, Industry-Wide Initiatives, and Democratic Governance Integration. Each recommendation addresses specific normative challenges while incorporating practitioner perspectives through direct quotes.

The methodology's unique contribution lies in bridging theoretical understanding with practical implementation guidance. By combining thematic analysis with theoretical frameworks and regulatory mapping, it provides construction industry stakeholders with actionable insights that balance technical requirements, ethical considerations, and organizational realities. This approach acknowledges AI literacy in construction PRM requires both empirical understanding of industry challenges and normative alignment with emerging regulatory landscapes.

Results

Theme 1: Regulatory and Governance Frameworks:

The implementation of AI in construction project risk management faces significant challenges regarding regulatory and governance frameworks. Industry professionals consistently highlight concerns about the absence of clear legal structures governing AI applications. As one participant noted,

"First of all, where is the legal side? Who is legally responsible if something goes wrong?" (Participant A9).

The characterization of AI systems as "black boxes" (Participant A9) further emphasizes the transparency concerns.

The integration of AI requires both technical competency and regulatory compliance, analogous to operating a vehicle. Additionally, AI's role in decision-making processes was acknowledged, with Participant A5 observing that

"...whilst you're making that decision, the effective data capture and cascading from that evaluation process can also help people to make informed decisions about resource management."

These findings indicate that while AI presents significant potential, the lack of established regulatory frameworks creates uncertainty in implementation, particularly in areas requiring human accountability.

Theme 2: Knowledge management and standardization

Construction project risk management (PRM) face significant operational challenges characterized by fragmented processes and inadequate formal protocols. Industry practitioners highlight the prevalence of manual methodologies, as evidenced by Participant A7's observation that

"...most of the risk management across industries including construction is very manual."

A critical concern emerges regarding the inefficient handling of institutional knowledge, with practitioners noting that

"One of the most significant risks in construction is knowledge management... people are consistently trying to solve problems that have already been solved."

This reflects the absence of systematic knowledge retention mechanisms and standardized methodologies. The current reliance on individual expertise rather than organizational learning systems impedes efficient risk assessment and management processes, suggesting a need for more structured approaches to knowledge management in construction PRM.

Theme 3: Socio-economic disparities:

The implementation of AI/ML in construction project risk management reveals significant socioeconomic disparities between organizations. Financial barriers, as noted by Participant A1:

"SMEs also want to incorporate but it's a balance of book SMEs won't have that kind of revenue," create fundamental adoption challenges for smaller firms. The expertise gap compounds these issues, with Participant A6 observing:

"One challenge for buying it is to have the expertise to using it... those that have the expertise get high jacked by big companies very quickly."

This creates a three-tiered system: large organizations with robust resources, mid-sized firms with moderate capabilities, and small organizations facing significant barriers. Recent regulatory frameworks, including the White House Executive Order on AI (2023) and EU AI Act (2024), inadvertently amplify these disparities through uniform requirements, suggesting a need for targeted interventions to prevent a two-tier industry division.

Theme 4: Ethical implementation and human oversight:

The adoption of AI in construction project risk management raises significant ethical considerations regarding responsible implementation and human oversight. Participant A6 emphasizes data responsibility, noting:

"If that data isn't used respectively and responsibly, then it can end up really... challenging human rights issues."

This concern highlights the critical importance of ethical data management practices.

The balance between AI capabilities and human judgment emerges as a key theme. As Participant A3 observes:

"I think the natural feature is that we will always want to double check we are humans... the machines are only as good as what we put in..."

While AI offers potential benefits, particularly in improving accessibility, as noted by Participant A1:

"ML can be used as a tool to inform the workforce about risks beforehand by handling different languages, making it easier for workers from diverse backgrounds to understand key information,"

successful implementation requires careful attention to workforce engagement and autonomy. This is reflected in the emphasis that "The important thing is to make sure that the top people are not pushing them and forcing them to do things. They must feel like they are part of the solution."

These perspectives underscore the need for an implementation approach that maintains human agency while leveraging AI's capabilities to enhance construction risk management processes while highlighting the normative challenges of AI literacy.

Discussion

The thematic analysis of professional perceptions across the construction industry showed that AI literacy is not simply a matter of technical competence or digital adoption rather it is fundamentally a normative challenge. That is, AI literacy in construction project risk management (PRM) is embedded within a landscape of ethical, legal, social, and institutional responsibilities many of which is unexplored or inadequately addressed by current regulations.

Theme 1: Regulatory and governance framework

The analysis of regulatory and governance framework revealed significant organizational challenges in implementing AI literacy to navigate complex institutional pressures within construction PRM. The participant's fundamental question about the legal accountability exposes what institutional theorists identify as "organizational uncertainty" in emerging technological domains (DiMaggio & Powell, 1983). This uncertainty is particularly critical in construction PRM where the consequences of AI decision-making can directly impact human safety, financial outcomes, and structural integrity.

Construction firms operating internationally face the challenge of developing governance frameworks that satisfy both the Executive Order's emphasis on "safe, secure, and trustworthy" AI development and Article 4's mandate for "sufficient level of AI literacy." The regulatory divergence forces organizations into structural changes driven by external pressures rather than efficiency considerations (Gillner, 2024)

Participant A9's observation about the "black box" nature of machine learning presents a critical governance challenge. Gunningham's (2012) reflexive regulation theory helps explain this tension: construction organizations must demonstrate regulatory compliance (reflexivity) while working with inherently opaque systems (technological limitation). Power (2007 raised a question on how organizations can prove compliance with transparency requirements when the technology itself resists transparency? (Power, 2007)

The car analogy provided by another participant offers insight into this contradiction. The participant intuitively understood that AI governance requires both operational competency ("know how to drive") and regulatory awareness ("follow all the rules of the road"). This dual requirement aligns with Gunningham's concept of "regulatory escalation" - the need for increasingly sophisticated governance mechanisms as technological complexity increases. However, unlike driving a car where rules and capabilities are well-established, AI governance in construction exists in what regulatory scholars call "institutional voids" - spaces where formal rules lag behind technological development (Doh et al., 2017)

Participant A5's insight about capturing, cascading and informed decision making about resource management reveals potential for what reflexive regulation. Drawing from Parker and Nielsen's (2011) compliance-plus framework, the data suggests that AI implementation can transcend basic regulatory compliance to generate broader organizational benefits in resource management and decision-making processes. This indicates that AI literacy standards may serve both regulatory and operational optimization purposes (De Almeida et al., 2021).

The governance frameworks emerging around AI literacy in construction PRM represent adaptive approaches to governing technologies whose impacts remain partially unknown (Gilad, 2010). As organizations effectively become regulatory subjects in learning which means that entities that must learn to regulate it technically as well. This dual development process creates better innovative oversight but also risks of regulatory capture or evasion.

Hardy and Maguire (2008) institutional entrepreneurship reveals that effective AI literacy governance in construction PRM requires creative responses to regulatory requirements that generate new organizational capabilities while maintaining compliance.

As organizations navigate these challenges, they are effectively creating new institutional logics that will shape the future of AI integration in construction industries globally (McKague, 2011)

Theme 2: Knowledge Management and Standardization

The results revealed construction PRM's persistent reliance on tacit knowledge, creating fundamental barriers to AI literacy integration. Participant A7's observation about manual risk management practices exemplifies the tendency for valuable knowledge to remain empirical and undocumented despite its critical importance to organizational success. (Nonaka and Takeuchi, 1995; Hoe, 2006)

This manual orientation reflects deep-rooted professional practices where risk assessment relies heavily on experienced practitioners' intuition and accumulated wisdom. The SECI model shows that construction organizations face AI integration challenges by remaining trapped in informal knowledge sharing practices between individuals. Without systematic approach to convert tacit knowledge into explicit, organizations cannot create the standardized knowledge bases necessary for ML training and implementation (Natek and Ledjak, 2021)

The regulatory requirement for "sufficient level of AI literacy" essentially demands that construction organizations start moving from tacit knowledge sharing through systematic documentation to integrated AI-enhanced decision-making systems. This represents a paradigmatic shift in how construction professionals conceptualize and manage risk knowledge.

The critical observation by participant A5 on knowledge fragmentation exposes systemic failures in the knowledge cycle. This redundancy represents barriers to effective knowledge transfer that create organizational incompetence (Watson, 2020)

From the SECI perspective, construction organizations appear trapped in repeated socialization cycles without progressing to externalization. Each project team rediscovers solutions rather than accessing and building upon documented knowledge from previous projects. This failure to complete the knowledge conversion process creates breaks in the continuous flow of knowledge transformation that characterize learning organizations.

Harfouche et al, 2023 argued the lack of systematic knowledge documentation which creates fundamental barriers to AI literacy in construction. AI systems require articulated knowledge for algorithmic processing, yet construction firms primarily operate in the "tacit dimension," creating an inherent incompatibility between current practices and AI implementation requirements (Addis, 2016)

Analysis reveals three key standardization gaps in construction AI literacy: nomenclature, processes, and outcomes. Drawing from Carlile's (2004) framework, these gaps represent knowledge barriers that conflict with traditional construction practices. The industry must develop standardized systems that balance knowledge translation while maintaining project-specific flexibility.

The knowledge management challenges revealed in this theme suggested that effective AI literacy in construction PRM requires more than technical training. Organizations must fundamentally transform their approach to knowledge creation,

capture, and distribution. The SECI model indicates that successful AI integration demands systematic approaches to converting experiential knowledge into reusable organizational assets. (Brown and Duguid, 2001)

This transformation has particular implications for meeting regulatory requirements. Article 4 of the EU AI Act's emphasis on "sufficient level of AI literacy" cannot be achieved through isolated training initiatives. Instead, organizations must develop systematically externalize tacit knowledge, combine explicit knowledge from multiple sources, and internalize standardized practices across project teams. Without fundamental changes to knowledge management practices, organizations risk investing in AI technology while lacking the necessary infrastructure to effectively implement and sustain it. This misalignment between technological adoption and knowledge management capabilities threatens the long-term success of AI literacy initiatives in construction.

The theme ultimately reveals that knowledge management standardization represents not merely a technical challenge but a fundamental prerequisite for the normative goals of AI literacy regulations - ensuring safe, responsible, and effective AI implementation in construction project risk management.

Theme 3: Socio-economic Inequities

Resource Dependence Theory explains how AI literacy regulations create systematic advantages for resource-rich construction firms. Participant A1's remarks on revenue difference explains structural barriers that transcend simple financial limitations to become existential compliance challenges (Pfeffer and Salancik, 2015). Regulatory requirements for "sufficient level of AI literacy" essentially create "compliance fixed costs" that disproportionately burden smaller organizations, unexpectedly undermining the democratic resolution of these frameworks.

Participant A6's observation that expertise gets "high jacked by big companies very quickly" reveals systematic labour market dynamics that compound resource inequalities. The uniform AI literacy requirements in both US and EU regulations may unintentionally increase industry inequality and may create a cyclical pattern where wealthy firms attract expertise, further increasing their advantages. The result is a three-tiered industry structure where large companies dominate AI capabilities, mid-sized firms remain vulnerable to talent loss, and SMEs face fundamental barriers to AI adoption which as a result reshape competitive landscapes around access to AI literacy capabilities (Markevicius, 2025; Phelps et al., 2007).

Resource inequities extend to individual workers, creating structural constraints on professional development independent of individual merit. This worker-level stratification challenges the human rights dimensions emphasized in AI regulations, suggesting that market-driven expertise concentration undermines efforts to democratize AI literacy across the construction workforce (Shuck et al., 2016).

Current AI regulations risk widening the gap between large and small construction firms. Drawing from institutional theory, standardization efforts may primarily benefit well-resourced organizations while creating barriers for smaller companies. To achieve equitable AI adoption, regulatory frameworks must incorporate flexible compliance pathways and support mechanisms for resource-constrained organizations, preventing industry consolidation and promoting democratic technological advancement (Peter, 2011; Shandilya et al., 2024).

Theme 4: Ethical Implementation and Human Oversight

Participant A6's concern about improper data usage aligns with Friedman et al.'s (2003) Value-Sensitive Design theory, highlighting how AI implementation in construction involves balancing competing ethical priorities. Organizations must consider worker privacy, safety, and efficiency while managing sensitive data, demonstrating that technical decisions carry significant implications for workplace rights beyond basic regulatory compliance which directly implicates both GDPR's data protection requirements and the EU AI Act's risk classification frameworks. Organizations must ensure AI literacy programs meeting Article 4 requirements without compromising worker privacy rights under GDPR Article 7 (Zaeem and Barber, 2020; Lorè et al., 2023; Edwards, 2021)

The resistance to top-down approaches and inclusivity challenges traditional construction hierarchies. This aligns with stakeholder theory of considering democratic participation in technical decision-making. Participant A1's observation about multilingual risk communication demonstrate technical capabilities that inherently embed social values about inclusion and workplace equity (Gould, 2002)

Participant A3's insistence on human verification reflects Shneiderman's (2020) "Human-Centered AI" principle. This represents ensuring practical rather than procedural oversight Dignum (2018). Effective AI literacy requires developing ethical reasoning capabilities alongside technical skills, enabling critical evaluation of AI-generated recommendations.

The findings reveal that ethical AI literacy implementation demands responsible innovation as illustrated by Stilgoe et al. (2020). Construction organizations must integrate professional ethics (safety, expertise), regulatory requirements (transparency, accountability), and inclusive workplace practices. This requires "ethical impact assessment" approaches (Mittelstadt et al., 2016) that systematically evaluate how technical choices affect diverse stakeholder groups, moving beyond compliance toward fundamental examination of how AI systems reshape workplace relationships and distribute agency in construction PRM contexts.

Recommendations:

Education Framework Development

Construction industry needs AI literacy education frameworks across multiple competency levels. Organizations should develop training programs following the SECI model, with three tiers: digital literacy for frontline workers, AI application skills for managers, and AI governance for executives.

Educational institutions should collaborate with industry bodies to establish specialized AI literacy programs within existing construction management programs.

As Participant A5 emphasizes, both sectors should work more closely:

"…embedding academics into industry so both streams could benefit from each other academics having benefit from real world experience and industry would benefit from genuine research into novel tech and ideas."

This integration addresses the critical timing gap noted:

"…sometimes the time lapse is too great from one being transferred to the other."

Such collaboration ensures educational frameworks remain aligned with both regulatory requirements under Article 4 of the EU AI Act and practical industry needs. Critical to this framework is addressing the knowledge fragmentation identified in Theme 2. Educational programs must include "knowledge externalization workshops" where experienced practitioners document tacit risk management knowledge in formats suitable for AI system training.

As Participant A3 highlights, effective comprehension is essential:

"for people to use ML, they need to be trained and to understand the process. If they don't...how would they interpret the message or the values or feedback from ML... the output from the ML and how it works."

This understanding becomes particularly crucial during "critical emergencies," ensuring AI literacy education addresses both routine operations and high-stakes decision-making scenarios required by regulatory frameworks.

Professional Practice Evaluation

Implementation needs evaluation mechanisms measuring both individual and organizational AI literacy. Based on Resource Dependence Theory, frameworks should adapt to different organizational capacities while maintaining standards. Professional bodies should create assessment protocols integrated with existing performance evaluations.

User accessibility must remain central to evaluation criteria. Participant A4 emphasizes that

"…software should be easily manageable and operatable by workforce... extremely user-friendly for people from diverse backgrounds... user friendly to your target audience."

This requirement aligns with both GDPR accessibility provisions and the EU AI Act's transparency obligations, ensuring evaluation frameworks measure practitioners' ability to make AI systems genuinely accessible across diverse construction teams.

Organizations should implement "AI literacy scorecards" measuring progress across technical competency, regulatory compliance, and ethical implementation. These scorecards enable benchmarking against industry standards while identifying specific development needs. Importantly, evaluation frameworks must address the socio-economic inequities identified in Theme 3, providing differentiated assessment pathways for organizations with varying resource capabilities. This ensures smaller firms aren't disadvantaged by evaluation requirements designed for larger enterprises.

Industry-Wide Initiatives

Collective action represents the most effective approach to addressing systemic challenges in AI literacy development. The construction industry should establish an "AI Literacy Association" bringing together major contractors, SMEs, educational institutions, and regulatory bodies. This consortium would develop shared infrastructure, including open-source AI training datasets specific to construction risks, standardized evaluation tools, and collaborative learning platforms.

Government-led pilot programs represent a crucial catalyst for industry-wide adoption. As Participant A4 suggests:

"…if the government were to do some trials on various sites of applying different software and working with contractors...trialing different things pilot prototypes which can then create an awareness for the industry provide case studies."

Such initiatives would provide practical demonstrations of AI literacy requirements under both the White House Executive Order and EU AI Act, creating tangible pathways for compliance while building industry confidence.

International coordination between US and EU industry bodies proves essential given regulatory divergence. The consortium should develop "regulatory translation frameworks" helping organizations navigate jurisdictional differences while maintaining consistent ethical standards. This includes creating industry-specific interpretive guidance for ambiguous regulatory requirements, reducing compliance uncertainty identified in Theme 1. Regular "AI Literacy Summits" would facilitate knowledge exchange, policy dialogue, and collaborative problem-solving across the global construction industry.

Democratic Governance Integration

Effective AI literacy implementation requires fundamentally democratic approaches to technology governance within construction organizations. Based on findings emphasizing worker participation, organizations should establish "AI Advisory Committees" with representation across hierarchical levels, trades, and cultural backgrounds. These committees would have substantive input on AI system selection, deployment, and evaluation, ensuring the "top people are not pushing them and forcing them to do things" as highlighted by participants.

Organizational commitment proves essential for sustainable implementation. Participant A5 identifies a key success factor:

"Anything a company really prioritizes they will invest in and if a company invest in a dedicated team whose sole purpose is to enhance existing business processes with ML, then that strategy is hard to beat, and impact could exponentially higher."

This insight suggests democratizing AI literacy requires dedicated resource allocation combined with participatory governance structures, ensuring both top-down commitment and bottom-up engagement.

Critical to democratic governance is ensuring AI literacy programs themselves are co-created with end-users. Worker councils should participate in curriculum development, ensuring training addresses practical needs rather than abstract compliance requirements. This bottom-up approach addresses the human oversight emphasis in Theme 4 while building organizational buy-in necessary for sustainable implementation. This aligns with both EU AI Act Article 4 requirements and the White House Executive Order's focus on "safe, secure, and trustworthy" AI development

Conclusion

In conclusion, this study examines the normative challenges of AI literacy in construction project risk management (PRM) through four key themes: regulatory compliance, knowledge management, socio-economic equity, and ethical implementation. The findings demonstrate that successful AI integration requires industry-wide transformation addressing four critical dimensions: establishing democratic governance frameworks,

standardizing knowledge management processes, mitigating resource-based inequities, and embedding ethical considerations throughout implementation.

Through thematic analysis of professional experiences, grounded in established theoretical frameworks and contemporary regulatory requirements, this research provides actionable recommendations spanning education framework development, professional practice evaluation, industry-wide initiatives, and democratic governance integration. The study's unique contribution lies in bridging empirical observations with normative analysis, offering construction stakeholders a comprehensive roadmap for implementing AI literacy that balances technical capabilities with ethical responsibilities.

The research illuminates how current regulatory frameworks, while establishing critical standards, inadvertently create stratified compliance challenges that particularly disadvantage smaller organizations. Moreover, the persistent reliance on tacit knowledge in construction PRM presents fundamental barriers to AI integration, requiring systematic organizational transformation beyond individual skill development. Most significantly, the findings emphasize that sustainable AI literacy development demands collaborative approaches that ensure worker participation while maintaining human oversight of AI-generated decisions.

As the construction industry navigates the rapid evolution of AI technologies and associated regulatory landscapes, this study provides essential guidance for stakeholders committed to responsible, inclusive, and effective AI implementation. The identified normative challenges and proposed recommendations offer practical pathways for construction organizations to develop AI literacy capabilities that not only meet regulatory requirements but also strengthen organizational resilience, promote equitable access to technological benefits, and maintain fundamental ethical standards in high-stakes construction project environments. Future success in AI integration will depend on the industry's collective commitment to addressing these multifaceted challenges through coordinated, democratic, and ethically grounded approaches.

Future Research Directions

This study's findings suggest several key directions for future research. Quantitative validation of the identified normative challenges would provide statistical verification across a broader industry sample. Development of measurement frameworks would establish standardized methods for assessing AI literacy in construction organizations. Cross-cultural studies could reveal how these challenges vary globally, while longitudinal research would evaluate the effectiveness of literacy programs over time. These research directions would strengthen our understanding of AI literacy's role in successful ML integration within construction project risk management. These findings suggest that addressing normative challenges in AI literacy is essential for successful ML integration in construction PRM, requiring a balanced approach to both technical and ethical considerations.

Limitations

As this abstract was taken from a dissertation completed by a master's student as part of their graduate studies, several limitations should be acknowledged. The relatively small sample size of nine professionals, while providing valuable insights,

may not fully represent the diverse perspectives across the global construction industry. Additionally, the study's focus on professionals with direct ML implementation experience potentially excludes viewpoints from those struggling with initial AI adoption, which could have provided valuable insights into fundamental literacy challenges. Furthermore, as the research was conducted during a period of rapidly evolving AI regulations and standards, some of the findings regarding regulatory compliance may require updates to reflect the latest legislative developments.

References

1. Abbasnejad, B., Nepal, M.P., Mirhosseini, S.A., Moud, H.I. and Ahankoob, A., 2021. Modelling the key enablers of organizational building information modelling (BIM) implementation: An interpretive structural modelling (ISM) approach. Journal of Information Technology in Construction, 26, pp.974-1008.
2. Addis, M., 2016. Tacit and explicit knowledge in construction management. Construction management and economics, 34(7-8), pp.439-445.
3. Barbosa, M.W., Vicente, A.D.L.C., Ladeira, M.B. and Oliveira, M.P.V.D., 2018. Managing supply chain resources with Big Data Analytics: a systematic review. International Journal of Logistics Research and Applications, 21(3), pp.177-200.
4. Barbosa, A.D.S., Bueno da Silva, L., Morioka, S.N., da Silva, J.M.N. and de Souza, V.F., 2023. Integrated management systems and organizational performance: A multidimensional perspective. Total Quality Management & Business Excellence, 34(11-12), pp.1469-1507.
5. Biden, J.R., 2023. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence.
6. Bilal, M., Oyedele, L.O., Qadir, J., Munir, K., Ajayi, S.O., Akinade, O.O., Owolabi, H.A., Alaka, H.A. and Pasha, M., 2016. Big Data in the construction industry: A review of present status, opportunities, and future trends. Advanced engineering informatics, 30(3), pp.500-521.
7. Burr, C. and Leslie, D., 2023. Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. AI and Ethics, 3(1), pp.73-98.
8. Butt, J., 2024. Analytical study of the world's first EU Artificial Intelligence (AI) Act. International Journal of Research Publication and Reviews, 5(3), pp.7343-7364.
9. Carlile, P.R., 2004. The dynamics of firm knowledge and competitive advantage: A framework and case example. Organization Science, 15, pp.555-568.
10. Cheng, J.Y., Devkota, A., Gheisari, M. and Jeelani, I., 2024. Integrating a Module on Artificial Intelligence Literacy in an Undergraduate Construction Management Course. Proceedings of 60th Annual Associated Schools, 5, pp.93-101.
11. Chiu, I.H.Y. and Lim, E.W., 2021. Managing Corporations' Risk in Adopting Artificial Intelligence: A Corporate Responsibility Paradigm. Wash. U. Global Stud. L. Rev., 20, p.347.
12. Chiarella, M.L., 2023. Digital Markets Act (DMA) and Digital Services Act (DSA): New Rules for the EU Digital Environment. Athens JL, 9, p.33.
13. Datta, S.D., Islam, M., Sobuz, M.H.R., Ahmed, S. and Kar, M., 2024. Artificial intelligence and machine learning applications in the project lifecycle of the construction industry: A comprehensive review. Heliyon.
14. De Almeida, P.G.R., dos Santos, C.D. and Farias, J.S., 2021. Artificial intelligence regulation: a framework for governance. Ethics and Information Technology, 23(3), pp.505-525.
15. Dignum, V., 2018. Ethics in artificial intelligence: introduction to the special issue. Ethics and Information Technology, 20(1), pp.1-3.
16. DiMaggio, P.J. and Powell, W.W., 1983. The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. American sociological review, 48(2),

17. Edwards, L., 2021. The EU AI Act: a summary of its significance and scope. Artificial Intelligence (the EU AI Act), 1, p.25.
18. Musch, S., Borrelli, M.C. and Kerrigan, C., 2024. Bridging compliance and innovation: A comparative analysis of the EU AI Act and GDPR for enhanced organisational strategy. Journal of Data Protection & Privacy, 7(1), pp.14-40.
19. Farahani, F.M., 2024. Challenges and Barriers to Implementing Collaborative Governance for Linking Education and Industry. Management Strategies and Engineering Sciences, 6(3), pp.164-173.
20. Fernandes, E., Holmes, W. and Zhgenti, S., 2024. Article 4 AI Literacy. The EU Artificial Intelligence (AI) Act: A Commentary, p.66.
21. Fernández, M.A., 2025. Risk Management System (Article 9). The EU regulation on Artificial Intelligence: A commentary, p.139.
22. Friedman, B., Kahn, P. and Borning, A., 2002. Value sensitive design: Theory and methods. University of Washington technical report, 2(8), pp.1-8.
23. Gann, D., 2001. Putting academic ideas into practice: technological progress and the absorptive capacity of construction organizations. Construction Management & Economics, 19(3), pp.321-330.
24. Gilad, S., 2010. It runs in the family: Meta□regulation and its siblings. Regulation & Governance, 4(4), pp.485-506.
25. Gillner, S., Blankart, K.E., Bourgeois, F.T., Stern, A.D. and Blankart, C.R., 2024. The challenges of regulatory pluralism. Health policy, 149, p.105164.
26. Gould, C.C., 2002. Does stakeholder theory require democratic management?. Business & Professional Ethics Journal, 21(1), pp.3-20.
27. Gunningham, N., 2012. Confronting the challenge of energy governance. Transnational Environmental Law, 1(1), pp.119-135.
28. Halder, A. and Batra, S., 2024. Navigating the ethical discourse in construction: A state-of-the-art review of relevant literature. Journal of Construction Engineering and Management, 150(3), p.03124001.
29. Hanna, M., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M. and Rashidi, H., 2024. Ethical and Bias considerations in artificial intelligence (AI)/machine learning. Modern Pathology, p.100686.
30. Hardy, C. and Maguire, S., 2008. Institutional entrepreneurship. In The Sage handbook of organizational institutionalism (pp. 198-217). SAGE Publications Ltd.
31. Harfouche, A., Quinio, B., Saba, M. and Saba, P.B., 2023. The Recursive Theory of Knowledge Augmentation: Integrating human intuition and knowledge in Artificial Intelligence to augment organizational knowledge. Information Systems Frontiers, 25(1), pp.55-70.
32. Ho, C.M.F. and Oladinrin, O.T., 2019. A paradigm shift in the implementation of ethics codes in construction organizations in Hong Kong: Towards an ethical behaviour. Science and Engineering Ethics, 25, pp.559-581.
33. Hoe, S.L., 2006. Tacit knowledge, Nonaka and Takeuchi SECI model and informal knowledge processes. International Journal of Organization Theory & Behavior, 9(4), pp.490-502.
34. Karatsareas, P., 2022. Semi-structured interviews. Research methods in language attitudes, pp.99-113.
35. Lloyd-Walker, B., Crawford, L. and French, E., 2018. Uncertainty as opportunity: the challenge of project based careers. International Journal of Managing Projects in Business, 11(4), pp.886-900.
36. Long, D. and Magerko, B., 2020, April. What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-16).

37. Lorè, F., Basile, P., Appice, A., de Gemmis, M., Malerba, D. and Semeraro, G., 2023. An AI framework to support decisions on GDPR compliance. Journal of Intelligent Information Systems, 61(2), pp.541-568.
38. Maitz, K., Fessl, A., Pammer-Schindler, V., Kaiser, R. and Lindstaedt, S., 2022. What do construction workers know about artificial intelligence? An exploratory case study in an Austrian SME. In Proceedings of Mensch und Computer 2022 (pp. 389-393).
39. Markevičius, V., 2025. Small and medium enterprises (smes) growth: the role of digital transformation and open innovation (Doctoral dissertation, Vilniaus universitetas.).
40. McKague, K., 2011. Dynamic capabilities of institutional entrepreneurship. Journal of Enterprising Communities: People and Places in the Global Economy, 5(1), pp.11-28.
41. Mittal, S., Khan, M.A., Romero, D. and Wuest, T., 2018. A critical review of smart manufacturing & Industry 4.0 maturity models: Implications for small and medium-sized enterprises (SMEs). Journal of manufacturing systems, 49, pp.194-214.
42. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L., 2016. The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), p.2053951716679679.
43. Mohamed, M.A.H., Al-Mhdawi, M.K.S., Ojiako, U., Dacre, N., Qazi, A. and Rahimian, F., 2025. Generative AI in construction risk management: a bibliometric analysis of the associated benefits and risks. Urbanization, Sustainability and Society, 2(1), pp.196-228.
44. Morandín-Ahuerma, F., 2023. IEEE: a global standard as an ethical AI initiative. Principios normativos para una ética de la inteligencia artificial (127-136).
45. Musch, S., Borrelli, M.C. and Kerrigan, C., 2024. Bridging compliance and innovation: A comparative analysis of the EU AI Act and GDPR for enhanced organisational strategy. Journal of Data Protection & Privacy, 7(1), pp.14-40.
46. Natek, S. and Lesjak, D., 2021. Knowledge management systems and tacit knowledge. International Journal of Innovation and Learning, 29(2), pp.166-180.
47. Nonaka, L., Takeuchi, H. and Umemoto, K., 1996. A theory of organizational knowledge creation. International journal of technology Management, 11(7-8), pp.833-845.
48. Ng, D.T.K., Leung, J.K.L., Chu, S.K.W. and Qiao, M.S., 2021. Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, p.100041.
49. Oladinrin, T.O. and Ho, C.M.F., 2014. Strategies for improving codes of ethics implementation in construction organizations. Project Management Journal, 45(5), pp.15-26.
50. Parker, C. and Nielsen, V.L. eds., 2011. Explaining compliance: Business responses to regulation. Edward Elgar Publishing.
51. Peters, B.G., 2011. Institutional theory. The SAGE handbook of governance, pp.78-90.
52. Pfeffer, J. and Salancik, G., 2015. External control of organizations—Resource dependence perspective. In Organizational behavior 2 (pp. 355-370). Routledge.
53. Phelps, R., Adams, R. and Bessant, J., 2007. Life cycles of growing organizations: A review with implications for knowledge and learning. International journal of management reviews, 9(1), pp.1-30.
54. Pinski, M., 2024. Artificial Intelligence Literacy-Conceptualization, Measurement, Enablement, and Its Impact on Individuals and Organizations.
55. Power, D.J., 2007. A brief history of decision support systems. DSSResources. com, 3.
56. Regona, M., Yigitcanlar, T., Xia, B. and Li, R.Y.M., 2022. Opportunities and adoption challenges of AI in the construction industry: A PRISMA review. Journal of open innovation: technology, market, and complexity, 8(1), p.45.
57. Regulation, P., 2018. General data protection regulation. Intouch, 25, pp.1-5.
58. RIVERA, E.J., Müller, C.C., Rahat, R. and ElZomor, M., 2024, June. Empowering Future Construction Professionals by Integrating Artificial Intelligence in Construction-Management Education and Fostering Industry Collaboration. In 2024 ASEE Annual Conference & Exposition.
59. Saka, A.B. and Chan, D.W., 2020. Profound barriers to building information modelling (BIM) adoption in construction small and medium-sized enterprises (SMEs) An interpretive structural modelling approach. Construction Innovation, 20(2), pp.261-284.

60. Secundo, G., Mele, G., Passiante, G. and Ligorio, A., 2024. How machine learning changes Project Risk Management: a structured literature review and insights for organizational innovation. European Journal of Innovation Management, 27(8), pp.2597-2622.
61. Shuck, B., Collins, J.C., Rocco, T.S. and Diaz, R., 2016. Deconstructing the privilege and power of employee engagement: Issues of inequality for management and human resource development. Human Resource Development Review, 15(2), pp.208-229.
62. Shandilya, S.K., Datta, A., Kartik, Y. and Nagar, A., 2024. Navigating the regulatory landscape. In Digital Resilience: Navigating Disruption and Safeguarding Data Privacy (pp. 127-240). Cham: Springer Nature Switzerland.
63. Shneiderman, B., 2020. Human-centered artificial intelligence: Three fresh ideas. AIS Transactions on Human-Computer Interaction, 12(3), pp.109-124.
64. Stilgoe, J., Owen, R. and Macnaghten, P., 2020. Developing a framework for responsible innovation. In The ethics of nanotechnology, geoengineering, and clean energy (pp. 347-359). Routledge.
65. Supranaviciene, U., Gedviliene, G. and Vaitkute, L., 2024. Relationship Between Technological Literacy and Interdisciplinary Competence Development for VET Learners in Construction Sector. Learning in the Age of AI: Towards Imaginative Futures, p.140.
66. Tadimalla, S.Y. and Maher, M.L., 2024. AI Literacy for All: Adjustable Interdisciplinary Socio-technical Curriculum. arXiv preprint arXiv:2409.10552.
67. Van Audenhove, L., Vermeire, L., Van den Broeck, W. and Demeulenaere, A., 2024. Data literacy in the new EU DigComp 2.2 framework how DigComp defines competences on artificial intelligence, internet of things and data. Information and Learning Sciences, 125(5/6), pp.406-436.
68. Waqar, A., 2024. Intelligent decision support systems in construction engineering: An artificial intelligence and machine learning approaches. Expert Systems with Applications, 249, p.123503.
69. Watson, J., 2020. Knowledge erosion and degradation: a single case-study of knowledge risks and barriers in a multi-business organisation (Doctoral dissertation, Victoria University).
70. Yao, D. and de Soto, B.G., 2022. A preliminary SWOT evaluation for the applications of ML to cyber risk analysis in the construction industry. In IOP Conference Series: Materials Science and Engineering (Vol. 1218, No. 1, p. 012017). IOP Publishing.
71. Zaeem, R.N. and Barber, K.S., 2020. The effect of the GDPR on privacy policies: Recent progress and future promise. ACM Transactions on Management Information Systems (TMIS), 12(1), pp.1-20.

GENAI AND PROPORTIONALITY: EUROPEAN AND PORTUGUESE ETHICAL-LEGAL FRAMEWORK

Filipa Pais D'Aguiar[1] [0000-0003-1614-1521], Fernanda Duarte[2] [0000-0003-2156-6946], Nuno Silva[3] [0000-0003-0157-0710]

Keywords: AI Literacy, Consciousness, Human Rights, Proportionality

Extended Abstract

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 the usage of AI systems are being explored, as previously presented at the "IX Coimbra International Conference on Human Rights: a transdisciplinary approach" paper, in 2024, as well as the work developed, since 2020, by the Research Group Project on "Human Rights in the Digital Era: e-person and privacy", at CEJEIA (Centre for Juridical, Economic, International and Environmental Studies), at Lusíada University, Lisbon, Portugal and in collaboration with NUPEJ (Núcleo de Estudos Judiciários) at UFF/ Universidade Federal Fluminense, Niterói, Brazil.

Nevertheless, 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 strengthening the thin balance between regulation, compliance and Human-AI sustainable development [1]. Therefore, the research problem considers: Firstly, how AI systems regulation complies with explainability, transparency and accountability demand as well as with a responsible, educated, conscious, ethical and trustworthy usage and dissemination of AI systems. As AI continues to evolve, the legal and ethical implications of autonomous decision-making become increasingly complex [2]. The question of accountability and liability in AI-driven actions - particularly in GenAI systems - raises fundamental concerns regarding algorithmic opacity, bias, and unintended consequences [3, 4]. As seen in recent discussions on the EU AI Act [5] [6], the challenge lies not only in defining high-risk AI applications but also in establishing mechanisms for algorithmic transparency and explainability [7]. Ensuring that AI systems comply with human rights standards requires multidisciplinary oversight, where regulators, ethicists, and technologists work together to develop a governance framework that fosters both innovation and public trust [8].

Secondly, how the definition of "high-risk" AI systems may impact Human-AI sustainable development. The challenges faced during the preparatory legal work and

1 CEJEIA, Lusíada University, Lisbon, Portugal
2 INCT-InEAC, Heidelberg, Federal Fluminense University, 3 COMEGI, Lusíada University, Lisbon, Portugal
3 COMEGI, Lusíada University, Lisbon, Portugal, 13001224@lis.ulusiada.pt

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 [9] (p. 14-22). 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 [5]. 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 [10, 11, 12], 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) (e.g. in Brazil [13] and in Chile [14]. Therefore, we look forward to identifying the 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 (v.g., the usage of AI in courts, by judges, in a support level) [15, 16].

Thirdly, ponder the proportionality principle framework, within the scope of the EU AI Act and the GDPR, considering special categories of personal data. The pondering mainly focuses on the proportionality principle dimensions practical meaning and range in the scope of the art. 5.º, n. º 1, par. h) of the EU AI Act [5] combined with the art. º 9.º of the GDPR [17], that address AI prohibited practices within the treatment of special categories of personal data such as biometric data. We look forward to understanding how the proportionality principle may potentially contribute to the actual and ultima ratio protection of biometric data usage, neuro-rights and protection from mental hacking within the scope of EU AI Act.

Thus, finally, the real challenge related to misuse or abuse of AI does seem to rest on finding the balance between the protection of Human Rights, the proportionality principle and the regulation of AI systems, with a particular focus on those systems (such as GenAI) that comprise sensitive data collection, usage and storage such as facial/biometric data and neural data. Consequently, ethical principles, digital bias, data privacy and copyrights, digital literacy and education, are subjects that need to be considered not only by algorithm stakeholders (both mathematicians and providers) and tech companies but also by governments and civil society in general [1]. Preliminary research results seem to evidence that AI literacy measures regarding the topics referred to above (both coming from private and public sectors), may contribute significantly to strengthening this thin balance between regulation, compliance and Human-AI sustainable development [18] (p.23-25). For instance, digital literacy measures developed within national digital strategies that achieve all activity sectors and education levels (v.g., integrating mandatory AI Syllabus from kindergarten to high school, Lifelong Learning strategies and programs on AI made available for vulnerable populations and specialised AI education programs for professionals from sensitive areas, such as, education, health, justice and security) seem to enshrine the so-longed potential balance between Human-AI sustainable development as they provide both an educated consciousness on

the usage of AI systems, ensuring that the last word and decision rely on and result from an educated and conscious Human decision which, ultimately, translates into a real and effective protection of Human Rights.

Although the focus of this research is on Portuguese and European Union ethical-legal framework solutions on all these issues [19], the dialogue and contributions from other ethical-legal orders (v.g. Brazil and Chile) may be referred to in an innovative open-minded and comparative law perspective. Hence, we have adopted comparative law methodology, particularly, the micro-comparison methodology, the hypothetical-deductive methodological approach and historical methodological procedure on global inequalities and digital justice.

References

1. UNESCO: Consultation Paper on AI Regulation: Emerging Approaches across the World. Paris, France, 2024. https://www.hkdca.com/wp-content/uploads/2024/08/consultation-paper-ai-regulation-unesco.pdf
2. Gellers, J.C. Artificial Intelligence, animal and environmental law. 1.ª Ed. London and New York: Routledge Taylor & Francis Group, 2021. Ebook.
3. Rogerson, S. Ethical Digital Technology in Practice. Boca Raton, London : CRC Press, Taylor & Francis Group, 2023.
4. 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
5. UE. Parlamento Europeu e Conselho. The Artificial Intelligence Act - Regulation (EU) 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng ; and https://www.artificial-intelligence-act.com/
6. Gstrein, O. J. & Haleem, N. & Zwitter, A. (2024). General-purpose AI regulation and the European Union AI Act. Internet Policy Review, 13(3). https://doi.org/10.14763/2024.3.1790
7. Innerarity, D.: Artificial Intelligence and Democracy; UNESCO and CLACSO: Paris, France, 2024. https://unesdoc.unesco.org/ark:/48223/pf0000389736´
8. Mollick, Ethan (2024) – Co-inteligência: viver e trabalhar com IA. Trad. Luís Gonçalves. 1.ª ed. Porto Editora: Porto. (Ideias de ler).
9. Pais D'Aguiar, F. (2024) – Direitos Humanos e Inteligência Artificial: principais dimensões jurídicas, éticas, sociais e culturais, no contexto europeu e transnacional. In NETO, Carlos et al. (coord.) (2024) – Mentes digitais : do zero ao infinito : Homenagem à Professora Fernanda Duarte. Pref. Guilhermo Gama. Editora GZ: RJ. 1-36.
10. Damásio, A. Sentir & Saber: a caminho da consciência. 1.ª ed. Temas e Debates, Círculo de Leitores: Lisboa, 2020.
11. Sigman, M. O poder das palavras: a arte de conversar. Trad. Ana Mendes. Temas e Debates. 1.ª ed. Bertrand Editora: Lisboa, 2023.
12. Rovelli, C. A ordem do tempo. Trad. Silvana Cobucci. 1.ªed., 4.ª Republicação. Penguin Random House: Lisboa, 2022.
13. BRASIL. Câmara dos Deputados. Projeto de Lei 522/22. Portal da Câmara dos Deputados. 2022. Disponível em: https://www.camara.leg.br/propostas-legislativas/2317524. https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2317524&fichaAmigavel=nao
14. CHILE. MinCiencia Ministerio de Ciencia, Tecnología, Conocimiento e Innovación (2021) – Ley 21383. Biblioteca del Congreso Nacional de Chile. https://www.bcn.cl/leychile/navegar?idNorma=1166983
15. Martinho, A. (2023) - Digitalização, automação e Inteligência Artificial nos Tribunais Portugueses. In a REVISTA (03; Cultural) 167-178. Lisboa: STJ.

16. Liu, J.Z. & Li, X. How do judges use large language models? Evidence from Shenzhen. In Journal of Legal Analysis (16)235–259. 2025. https://doi.org/10.1093/jla/laae009
17. Parlamento Europeu e Conselho. Regulamento (EU) 2016/679, de 27 de abril de 2016 (Regulamento Geral sobre a Proteção de Dados). https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:32016R0679&from=EN.
18. Mazzi, F., Taddeo, M. & Floridi, L. AI in Support of the SDGs: Six Recurring Challenges and Related Opportunities Identified Through Use Cases. In Mazzi, Francesca and Floridi, Luciano, The Ethics of Artificial Intelligence for the Sustainable Development Goals (May 15, 2023). Philosophical Studies Series (PSSP, volume 152), 2023. https://ssrn.com/abstract=4448859
19. 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.
20. Government of Portugal, Lisbon (2024). ISSN (online): 1647-6212. https://www.gee.gov.pt/RePEc/WorkingPapers/GEE_PAPERS_186.pdf.

ARTIFICIAL INTELLIGENCE, NEURODIVERGENCE, AND ETHICAL DECISION-MAKING: THE NEED FOR INCLUSIVE FRAMEWORKS AND HUMAN OVERSIGHT

Elisabet Bolarín Miró[1] [0009-0001-0240-3888], M. Mercedes Carmona Martínez[2] [0009-0005-1987-3201], Asunción López Arranz[3] [0000-0002-5761-771X]

Keywords: Artificial Intelligence (AI), Neurodivergence, AI Ethics, Algorithmic Fairness, Legal and Ethical Implications.

Abstract

The rapid evolution of Artificial Intelligence (AI) is redefining decision-making in key areas such as employment, human resources, and corporate governance. However, as AI systems become more autonomous, ethical concerns about inclusivity have become increasingly prominent. One pressing yet often overlooked issue is the underrepresentation of neurodivergent individuals in AI development and governance. Neurodivergence [including autism, ADHD, and dyslexia], represents cognitive variations that fall outside conventional neurotypical frameworks. As a result, AI systems [designed primarily by neurotypical developers] often reinforce biases that marginalize neurodivergent populations.While accessibility initiatives exist for digital inclusion, the intersection between AI ethics and neurodivergence remains underexplored. This is concerning as AI increasingly shapes hiring and judicial decisions, where algorithmic biases can lead to systemic discrimination.Existing regulatory frameworks, such as the GDPR and Directive 2000/78/EC, uphold non-discrimination principles but fail to explicitly ensure fair neurodivergent representation in AI systems. Likewise, emerging mechanisms [such as the AI Regulation (RIIA) and the IAACT certification initiative] lack explicit provisions treating cognitive diversity as a foundational ethical component.This paper proposes an inclusive framework grounded in five pillars: participatory design, interdisciplinary oversight, regulatory adaptation, specialized training, and data representativeness. We argue that true technological ethics requires continuous human oversight and regulation that reflects cognitive diversity.

Introduction

The rapid evolution of Artificial Intelligence (AI) has redefined decision-making across industries, from healthcare and education to legal and corporate

1 Catholic University of San Antonio of Murcia (UCAM), Murcia, Spain, ebolarin@alu.ucam.edu
2 University of A Coruña, A Coruña, Spain
3 Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany, lncs@springer.com

However, as AI systems become more autonomous, ethical concerns regarding inclusivity and fairness have intensified. One of the most pressing but often overlooked issues is the underrepresentation of neurodivergent individuals in AI development and governance. Neurodivergence, encompassing conditions such as autism, ADHD, and dyslexia, represents cognitive variations that do not fit within conventional neurotypical frameworks. Consequently, AI systems [designed primarily by neurotypical developers] often reinforce biases that marginalize neurodivergent populations.

Sartor (2005) provides a legal-philosophical foundation for integrating cognitive diversity into normative systems, emphasizing the importance of diverse reasoning models in decision-making frameworks. Floridi and Cowls (2019) similarly argue that existing AI ethics frameworks rarely consider neurodivergent perspectives. While accessibility initiatives exist for disability inclusion in digital environments (Angelino et al., 2023), the intersection between AI ethics and neurodivergence remains underexplored. This oversight is concerning given that AI influences domains education, hiring processes, medical diagnoses, and criminal justice [where algorithmic biases ~~can~~ lead to systematic discrimination].

Current regulatory frameworks, including the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679) and Directive 2000/78/EC, establish principles of non-discrimination, but they do not explicitly ensure the fair representation of neurodivergent individuals in AI-driven decision-making. Similarly, emerging AI governance mechanisms [such as the Artificial Intelligence Regulation (RIIA, Royal Decree 2021/1234) and the Artificial Intelligence Technology Accreditation and Certification Initiative (IAACT, Resolution 2022/5678)] lack provisions that address cognitive diversity as a fundamental component of AI ethics.To address this gap, we propose a participatory, interdisciplinary framework that incorporates neurodivergent perspectives throughout the AI lifecycle. It is built on five core pillars: 1)Inclusive Co-Design: Involving neurodivergent individuals in system development to ensure representation. 2)Ethical Oversight Committees: Forming multidisciplinary teams with ethicists, legal experts, technologists, and neurodivergent representatives. 3) Dynamic Regulation: Updating AI laws to ensure transparency, accountability, and cognitive inclusion. 4)Ethics Education: Training policymakers, developers, and legal professionals in inclusive AI ethics. 4) Representative Data: Using training datasets that reflect the full spectrum of human cognition.

This research contributes to ongoing discussions in AI ethics by highlighting the necessity of neurodivergent-inclusive policies. True technological ethics lies in the ability to dynamically respond to human diversity, which can only be achieved through continuous regulatory adaptation and human oversight.

Main Objectives

Our objectives are to:

1. Critically assess current ethical and legal frameworks in AI, focusing on their failure to account for neurodivergent perspectives and the structural biases this omission generates.

2. Develop an inclusive ethical and legal model, integrating neurodivergence into AI governance, emphasizing human oversight, and proposing regulatory updates to ensure cognitive diversity is addressed.
3. Promote participatory design, actively involving neurodivergent communities in the creation and evaluation of AI technologies through inclusive methodologies and interdisciplinary collaboration.

Method

This study employs a qualitative methodology based on a systematic literature review and comparative analysis of interdisciplinary texts. The process was structured in three phases:

Source Selection and Review: Relevant academic literature was identified through searches in databases such as Google Scholar, Scopus, and Web of Science using keywords like "artificial intelligence," "neurodivergence," "inclusive ethics," and "human oversight." Key theoretical works (Sartor, 2005; Floridi & Cowls, 2019) and regulatory texts (European Commission, 2019) were selected.

- Source Selection and Review: Relevant academic literature was identified through searches in databases such as Google Scholar, Scopus, and Web of Science using keywords like "artificial intelligence," "neurodivergence," "inclusive ethics," and "human oversight." Key theoretical works (Sartor, 2005; Floridi & Cowls, 2019) and regulatory texts (European Commission, 2019) were selected
- Comparative Analysis: Ethical and legal frameworks were analyzed in relation to their responsiveness to neurodivergent needs. This included contrasting disability inclusion studies (Angelino et al., 2023; Berrío Zapata et al., 2020) with literature on algorithmic bias (Mittelstadt et al., 2016; Jobin, Ienca, & Vayena, 2019) and human oversight (Vayena, Blasimme, & Cohen, 2018).
- Famework Development: Insights were synthesized to formulate an inclusive ethical-legal model that integrates neurodivergent perspectives across the AI lifecycle, emphasizing the critical role of human intervention and participatory governance.

Results and Discussion

This study reveals that current ethical guidelines (European Commission, 2019; Jobin et al., 2019) follow a homogeneous approach that fails to accommodate cognitive diversity. Sartor (2005) underscores how regulatory reasoning often excludes neurodivergent perspectives, contributing to the systemic marginalization of individuals with atypical cognitive profiles.

Further, research by Mittelstadt et al. (2016) and Floridi & Cowls (2019) demonstrates that data and design biases in AI systems perpetuate social inequalities. In the case of neurodivergence, insufficient representation in training datasets amplifies this issue, leading to inadequate or even harmful technological outputs, particularly

in education, employment, health, and justice. These gaps underscore the necessity of human oversight to intervene and correct errors that algorithms cannot anticipate.Vayena, Blasimme, and Cohen (2018) emphasize the ethical importance of human control in automated decision-making, highlighting the risks of accountability gaps when oversight is absent. From a legal standpoint, existing regulations such as the GDPR (Regulation (EU) 2016/679) and Directive 2000/78/EC uphold principles of non-discrimination, yet lack mandates for ensuring cognitive diversity in AI training data. National-level initiatives such as the RIIA (Royal Decree 2021/1234) and IAACT (Resolution 2022/5678) promote ethical governance, but still fall short in addressing neurodivergent needs explicitly.

In response, this study proposes an inclusive governance framework built on five pillars:

- Participatory Design: Involving neurodivergent communities early in design and testing phases (Shinohara & Wobbrock, 2011).
- Ethical Oversight Committees: Multidisciplinary bodies including experts in ethics, law, medicine, AI, and neurodivergent representation.
- Dynamic Regulation: Updating legal norms to enforce transparency, accountability, and responsiveness to cognitive diversity.
- Training and Education: Equipping developers and policymakers with the tools to design inclusively.
- Representative Data: Ensuring training datasets reflect a full spectrum of cognitive variability through partnerships with research and advocacy groups.

Ethically, the framework aims to reduce systemic discrimination by embedding neurodivergent perspectives throughout the AI lifecycle. Interdisciplinary oversight ensures real-time error correction and safeguards individual rights. Furthermore, adaptive regulation allows AI to evolve in alignment with the complexity of human cognition[recognizing exceptions and nuances traditional systems often ignore].

Conclusions

This study highlights the urgent need to integrate cognitive diversity into the ethical and legal governance of Artificial Intelligence. While existing regulations such as the GDPR, Directive 2000/78/EC, RIIA, and IAACT promote principles of protection and fairness, they still fail to include the specific needs of neurodivergent individuals. This gap leads to biased systems that can reinforce social inequalities. The proposed framework offers a practical solution by embedding inclusion at every stage of the AI lifecycle. It promotes participatory design, human oversight, and the use of representative data to ensure fair outcomes for all users. Ensuring that technology works for everyone [especially those historically excluded [requires more than good intentions]. It requires strong laws, ethical awareness, and interdisciplinary cooperation. Only by adapting our systems to the reality of human diversity can we guarantee that AI supports justice, inclusion, and dignity for all.

DECISION-MAKING ABILITY: RECOMMENDATION ENGINE EXPOSITION

ANTONIO J. SEGURA[1] [0009-0003-3441-5079], BELÉN ZAPATA-BARRERO[2] [0000-0002-0786-5504]

Keywords: Decision-making, Recommendation engines, Automation and human agency, Negative feedback loop

Abstract

Recommendation engines have been scarce in recent decades in our digital interactions; today, they are the primary driver of many relevant interfaces, such as social networks. This paper shows the necessity of measuring and analyzing counterfactually the decision-making ability of the general population and discusses the relevance and limitations of this analysis. Many methodologies for measuring decision-making capability have been developed for analyzing high-performance e-Sport players; bringing these techniques to the general population reveals two problems: a Lack of common context and the exposition of multiple factors. As a foundational model, we propose a timed CAPTCHA-like test to assess decision-making ability, accompanied by a student survey on the time spent on social networks and video games, to measure differences in the population's habits. These studies should be extended in time for monitoring the improvement or degradation of the same population. This information will help us understand human cognitive adaptation and raise awareness of the negative externality associated with recommendation engines.

Introduction

The aim of this study is to highlight the potential risks associated with the degradation of human decision-making abilities. The central hypothesis suggests the existence of a negative feedback loop: as recommendation engines become more widely used, individuals make fewer decisions. Consequently, the decline in decision-making frequency weakens the cognitive processes required for making choices, further increasing reliance on recommendation engines to avoid the cognitive cost of decision-making.

Recommendation engine algorithms first emerged in the 1990s [7] and have since become integral to the most popular websites. Over the past few decades, these algorithms have been incorporated into streaming platforms, marketplaces, and social networks [12], [9]. For example, Amazon began integrating recommendation engines

1COMPLUTENSE UNIVERSITY OF MADRID, SPAIN, ANTOSEGU@UCM.ES
2UNIVERSITY OF MURCIA, SPAIN, BELEN.ZAPATA@UM.ES

into its user interface in the early 2000s [4] and continues to enhance these systems, with an update to its recommendation features published last year [1].

As a result, a portion of human decision-making is increasingly outsourced to algorithms. Given that decision-making is a trainable skill [10], [11], a reduction in its use may lead to its deterioration, reinforcing the negative feedback loop.

Certain activities, such as content creation and video gaming, may counteract the effects of recommendation engines by exposing users to environments that require frequent decision-making. However, these activities do not impact all individuals equally, leaving some populations at greater risk of reduced exposure to decision-making processes.

In this study, we propose and discuss the limitations of an experimental framework designed to test our hypothesis and identify population groups most affected by this phenomenon. Finally, we outline additional experimental data required to gain a deeper understanding of decision-making degradation.

Decision-making ability experimental state of the art

Decision-making *ability* seems to be trained, studies like [14] shows differences between beginners and experts video-game players. There are many proposed method for measure decision ability in professional video-game at [5]. Qui et al. [6] propose a method based in a test for recognition of different states of game given an image.

Since decision-making ability is trainable, a lack of basal training due to a reduction in the number of decisions made by an individual generates a relative reduction of decision-making ability, at least in comparison to other individuals. This hypothesis implies that subrogating decisions to recommendation engines reduces the number of decisions actually made. We define decision subrogation as the delegation of any decision that used to be chosen by an individual to an algorithm. There is currently no evidence for testing this hypothesis and its implications. One of the objectives of this paper is to propose an experimental framework to accept or reject this hypothesis.

Methodology

The *proposed methodology has two requirements*. First, the method has to be applied to the general population, which presents some issues because there is not a common knowledge background, therefore there could be biases in the measurement. The second requirement is to measure the positive and negative exposition factors for each individual. To quantify the impact of these factors, we propose a counter-factual analysis after measurement.

We propose, a visual method similar to [6] for decision-making ability. In order to adapt this test to the general population, our proposal is similar to a CAPTCHA. It shows daily life pictures and asks questions about them. For example: Given a picture showing a street with a lot of people, one of them seems to need medical help and another seems to be a doctor, the image is shown to the subject during 1 second. After watching the image, the subject has to answer what they would do. We evaluate answers in terms of elements detected in the picture. For this example, we propose a three-point

scale: One, if the health risk situation is not detected. Two, if the health risk situation is reported but the doctor is not detected. Finally three, if both the health risk situation and doctor are detected.

For monitoring detected exposition factors like video-games and social networks exposure, we propose a survey conducted before the CAPTCHA test. A counterfactual analysis is required in order to separate two opposite detected factors: video-game exposition as a positive feedback factor and recommendation engine exposition as a negative feedback factor.

As counter-factual method we propose a Propensity Score Matching (PSM) [8]. Given the standardized test result as the target variable and two explanatory variables from survey, recommendation engine exposition (RE) and video-games exposition (VG).

Results

According to the bibliography, decision-making can be trained [14]. This fact implies that people who are exposed to a positive feedback factor like "videogames" could show more capacity for decision-making than people exposed to a negative feedback factor like recommendation engine decision subrogation.

Recommendation engines as choice aggregators generate perverse incentives in terms of public choice strategies. If they produce a deterioration of decision ability, we could find a negative feedback loop that generates a dependency on these recommendation engines. Since aggregation choice methods have to be agreed by other aggregation choice methods, another negative feedback loop is generated that strengthens the dependency on these systems.

Conclusions

New evidence is required in order to understand decision-making ability in the general population and its development regarding habits. In this paper, we propose an experimental framework that should be implemented. New factors that affect decision-making ability should be analyzed and included in the experiments.

Deterioration of decision-making abilities presents two levels of negative feedback regarding recommendation engine exposure. The first negative feedback occurs because more exposure to recommendation engines implies less decision making ability is developed. The second negative feedback process is produced because recommendation engines are also choice aggregators, and since a choice aggregator is a network good [2], once it is established as a choice aggregator method, it is very difficult to return to another method. Due to these two reasons, recommendation engine exposure could produce a rapid loss of decision-making ability.

Limitations and future research

The Proposed methodology shows some limitations. More relevant factors should be detected, and in order to measure degradation, we need longitudinal time studies. *Future research* should implement the proposed experiments and study more possible factors.

Relevance

Public goods , according to Stiglitz definition [13]: "Nonrivalrous consumption—the consumption of one individual does not detract from that of another— and nonexcludability—it is difficult if not impossible to exclude an individual from enjoying the good" play a main role in current society [3]. To keep a good decision-making ability in the general population allows for a correct alignment in management of public goods.

Decision-Making Ability: Recommendation Engine Exposition

Acknowledgments: We would like to acknowledge the reviewers and interested parties in the proposed discussion.

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

References

1. Cheedella, K., Fathimabi, S., Chinamuttevi, D.: Amazon product recommendation system using apache spark. In: 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence). pp. 223–227. IEEE (2024)
2. Katz, M.L., Shapiro, C.: Systems competition and network effects. Journal of economic perspectives 8(2), 93–115 (1994)
3. Kaul, I., Grunberg, I., Stern, M.: Global public goods. Global public goods 450 (1999)
4. Leavitt, N.: Recommendation technology: Will it boost e-commerce? Computer, 39(5), 13–16 (2006)
5. Pedraza-Ramirez, I., Musculus, L., Raab, M., Laborde, S.: Setting the scientificstage for esports psychology: A systematic review. International Review of Sport and Exercise Psychology 13(1), 319–352 (2020)
6. Qiu, N., Ma, W., Fan, X., Zhang, Y., Li, Y., Yan, Y., Zhou, Z., Li, F., Gong, D.,Yao, D.: Rapid improvement in visual selective attention related to action video gaming experience. Frontiers in human neuroscience 12, 47 (2018)
7. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: An openarchitecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM conference on Computer supported cooperative work. pp. 175–186 (1994)
8. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika 70(1), 41–55 (1983) 9. Schrage, M.: Recommendation engines. MIT press (2020)
10. Scopelliti, I.: Training can improve decision making. The science of beliefs: A multidisciplinary approach pp. 557–573 (2022)
11. Sherstobitova, A.A., Kazieva, B.V., Glukhova, L.V., Kaziev, V.M., Koroleva, E.I.: Decision-making training for students and managers using data science and smart platforms. In: International KES Conference on Smart Education and Smart ELearning. pp. 163–172. Springer (2023)
12. Singhal, S., Pal, K.: State of art and emerging trends on group recommender system: a comprehensive review. International Journal of Multimedia Information Retrieval 13(2), 22 (2024)
13. Stiglitz, J.E., et al.: Knowledge as a global public good. Global public goods: International cooperation in the 21st century 308, 308–325 (1999)
14. Valls-Serrano, C., de Francisco, C., Caballero-López, E., Caracuel, A.: Cognitiveflexibility and decision making predicts expertise in the moba esport, league of legends. Sage Open 12(4), 21582440221142728 (2022)

ARTIFICIAL INTELLIGENCE: INTERPRETABILITY AND MEASUREMENT OF ITS UNDERLYING ETHICS

Miguel Houghton Torralba[1], Ziwei Shu[2] [0000-0001-6788-1111],
Ramón Alberto Carrasco[3] [0000-0001-7365-349X],
María Francisca Blasco López[4] [0000-0002-6660-3571]

Keywords: Artificial Intelligence, Corporate Social Responsibility, Measurement, Interpretability, Ethics.

Extended Abstract

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. The idea that AI must be interpretable and explainable has become a key theme in public and academic discourse regarding the ethical implications of this rapidly evolving technology. A primary driver of this concern is the widespread use of neural networks, which are often described as "black box" models due to their opaque internal reasoning processes, despite their capacity to perform complex tasks with high accuracy [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.

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. On the other hand, the concept of Measurement of underlying ethics involves assessing how ethical considerations are integrated into AI systems. This includes evaluating "intangible" characteristics like transparency, fairness, and accuracy. Therefore, it should be essential to develop frameworks and metrics that can help organizations determine whether their AI systems align with ethical standards and societal values.

1 Department of Marketing, Faculty of Commerce and Tourism, Complutense University of Madrid, Madrid 28003, Spain
2 Department of Marketing, Faculty of Statistics, Complutense University of Madrid, Avenida Puerta de Hierro,1, 28040 Madrid, Spain, ziweishu@ucm.es
3 Department of Marketing, Faculty of Statistics, Complutense University of Madrid, Avenida Puerta de Hierro,1, 28040 Madrid, Spain
4 Department of Marketing, Faculty of Commerce and Tourism, Complutense University of Madrid, Madrid 28003, Spain

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.

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. 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.

Furthermore, it is important to consider the concept of corporate social responsibility (CSR), which can be considered a critical element of modern business practices, ensuring that companies operate in a manner that benefits society while maintaining integrity. CSR and AI constitute a binomial that encompasses a broad range of activities, including environmental sustainability, ethical labor practices, philanthropy, and community engagement [6]. 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 [7]. Ethical business conduct, on the other hand, involves adherence to moral principles, transparency, and compliance with legal standards [8].

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 [9]. 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 these "advices" 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 [10]. Additionally, organizations engage in philanthropic activities by donating to charitable causes and supporting community development initiatives.

Ethical decision-making in business is guided by principles such as honesty, integrity, and accountability. Companies that uphold ethical standards foster trust among stakeholders, including customers, employees, and investors [11]. Unethical practices, such as corruption, fraud, or exploitative labor, can damage a company's reputation and lead to legal consequences. As a result of this, the underlying ethics over AI that manage our abstract should be seriously considered.

Evolving in these concepts of CSR with AI, it is logic to admit that its Interpretability, Measurements and ethics should be bonded. AI is transforming industries, but its development and ways to measure raises significant ethical concerns and CSR obligations. Firms utilizing AI must balance innovation with ethical considerations, ensuring fairness, transparency, and accountability [12]. 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 [13]. 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 [14].

Another CSR concern is job displacement. It is a general concern that probably millions of current jobs are at risk, requiring businesses to invest in workforce reskilling and education initiatives [15]. Additionally, AI's environmental impact, particularly in energy-intensive machine learning models, calls for sustainable AI development practices that should also be assessed carefully [16]. 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.

In addition, in the current environment where customers' opinions are of fundamental importance, one of the major challenges companies faces, 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 [17]. The ability to accurately assess this gap depends heavily on the reliability of the data collected, which must adhere to established service quality standards. Consequently, the effectiveness of such evaluations is often conditioned by the robustness of the survey instruments and the interpretability of the data.

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 [17] 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. [19]. Furthermore, the work of Carrasco et al. [20] 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.

To sum up, the development and interpretation of metrics in AI are essential for ensuring the effective application of CSR in increasing technological societies. 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 transparency, accountability, and interpretability 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.

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. Technological Forecasting and Social Change, 174, 121209 (2022)
3. Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S.: Transparency and explainability of AI systems: From ethical guidelines to requirements. Information and Software Technology, 159, 107197 (2023).
4. Vainio-Pekka, H., Agbese, M. O.-O., Jantunen, M., Vakkuri, V., Mikkonen, T., Rousi, R., & Abrahamsson, P.: 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, 1–17 (2024).
6. Carroll, A. B.: The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. Business Horizons, 34(4), 39–48 (1991).
7. Kotler, P., & Lee, N.: Corporate Social Responsibility: Doing the Most Good for Company and Your Cause. Wiley (2005).
8. Crane, A., Matten, D., Glozer, S., & Spence, L.: Business Ethics: Managing Corporate Citizenship and Sustainability in the Age of Globalization (5th edition). Oxford University Press (2019).
9. Kramer, M. R., & Porter, M.: Creating shared value. Harvard Business Review (2011).
10. Parmar, B. L., Freeman, R. E., Harrison, J. S., Wicks, A. C., Purnell, L., & De Colle, S. : Stakeholder theory: The state of the art. The academy of management annals, 4(1), 403-445 (2010).
11. Trevino, L. K., & Nelson, K. A.: Managing Business Ethics: Straight Talk about How to Do It Right. John Wiley & Sons (2016).
12. Floridi, L., & Cowls, J.: A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1) (2019).
13. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (2016).
14. Doshi-Velez, F., & Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning (arXiv:1702.08608) (2017).

15. Brynjolfsson, E.: The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies (2014).
16. Strubell, E., Ganesh, A., & McCallum, A.: Energy and Policy Considerations for Modern Deep Learning Research. Proceedings of the AAAI Conference on Artificial Intelligence, 34(09), 13693-13696 (2020).
17. Chumpitaz Caceres, R., & Paparoidamis, N. G.: Service quality, relationship satisfaction, trust, commitment and business- to-business loyalty. European Journal of Marketing, 41(7/8), 836–867 (2007).
18. Herrera, F., & Martinez, L.: A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Transactions on Fuzzy Systems, 8(6), 746–752. IEEE Transactions on Fuzzy Systems (2000).
19. Herrera, F., Martınez, L., & Sánchez, P. J.: Managing non-homogeneous information in group decision making. European Journal of Operational Research, 166(1), 115–132 (2005).
20. Carrasco, R., López, F., Garcia-Madariaga, J., & Herrera-Viedma, E.: A Fuzzy Linguistic RFM Model Applied to Campaign Management. International Journal of Interactive Multimedia and Artificial Intelligence (2018).

ETHICAL PRINCIPLES FOR THE DEVELOPMENT OF OFFICIAL STATISTICS USING MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE TECHNIQUES

Jorge-Eusebio Velasco-López[1], Ramón-Alberto Carrasco[2], Gema Fernández-Avilés[3], José-María Mon- tero-Lorenzo[4]

Keywords: Ethical principles, Official Statistics, Artificial Intelligence.

Abstract

The development and implementation of Artificial Intelligence (AI) and Machine Learning (ML) systems in the field of official statistics require an ethical and regulatory approach to ensure the protection of human rights, privacy, and fairness. The 2024 Artificial Intelligence Strategy emphasizes the importance of re- sponsible AI, ensuring 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.

Introduction

At the international level, various institutions have developed ethical principles applicable to the use of AI in the statistical context. UNESCO, through the Recommendation on the Ethics of Artificial Intelligence [2], and the OECD, with its AI Principles [3], have outlined guidelines for the ethical development and application of these systems. In the European Union, the High-Level Expert Group on AI has defined guide- lines for trustworthy AI based on principles of transparency, fairness, and human oversight [4].

National Framework

At the national level, Spain has integrated ethical principles into the production of official statistics, aligned with Regulation (EC) No. 223/2009 [5] and the European Statistics Code of Practice [6] (Eurostat, 2022). These principles include professional independence, impartiality, objectivity, reliability, statistical confi- dentiality, transparency, and proportionality. The application of ML-AI to official statistics must adhere to these principles to ensure the quality and reliability of the produced information.

1Instituto Nacional de Estadística; jorge.velasco.lopez@ine.es
2 Department of Marketing, Faculty of Statistics, Universidad Complutense de Madrid; ramoncar@ucm.es
3Faculty of Legal and Social Sciences, University of Castilla-La Mancha; Gema.FAviles@uclm.es
4Faculty of Legal and Social Sciences, University of Castilla-La Mancha; Jose.MLorenzo@uclm.es

Specific Ethical Principles for AI in Official Statistics

Building upon general ethical principles, specific considerations like explainability, bias mitigation, and data protection are critical for applying AI in official statistics. Recommendations from UNESCO and OECD emphasize transparency, fairness, and robust governance to prevent discriminatory biases and en- sure respect for human rights [2, 3].

To address technical challenges such as explainability and data quality, statistical agencies must invest in computational infrastructure and algorithmic transparency while engaging stakeholders to foster trust. Stakeholder input is crucial to aligning AI applications with societal values, particularly in areas where biases or misinterpretation could compromise public confidence in statistical outputs.

Governance and Oversight Framework

Ethical models proposed by international organizations agree on the necessity of establishing a robust gov- ernance 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 discrim- inatory biases and ensuring equitable access to information.

The 2019 Ethics Guidelines for Trustworthy AI outline seven non-binding ethical principles. These guidelines aim to promote trustworthy AI supported by three essential components throughout the system's lifecycle: (a) AI must be lawful, complying with all applicable laws and regulations; (b) it must be ethical, ensuring respect for ethical principles and values; and (c) it must be robust, both technically and socially, as AI systems, even with good intentions, can cause unintended harm.

1. Respect for human autonomy
2. Prevention of harm
3. Fairness
4. Explainability

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.

Implementation in Official Statistics

Applying these principles to official statistics requires the development of specific codes of conduct for the use of ML-AI, aligned with official statistical principles and international guidelines. These codes should establish clear standards for transparency, fairness, and data protection, ensuring that AI applications in official statistics are conducted 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 produc- tion of official statistics is governed by Article 338.1 of the Treaty on the Functioning of the European Union, Regulation 223/2009, and the European Statistics Code of Practice (ESCoP), ensuring fundamental principles such as independence, impartiality, reliability, transparency, and proportionality.

Empirical examples, such as the application of deep learning to automate data collection processes or the use of synthetic data to address privacy concerns, will illustrate the transformative potential of AI in statistical agencies. Forward-looking strategies should also explore how AI automation could enhance ef- ficiency while safeguarding ethical standards.

Conclusion

The use of AI-ML in official statistics must adhere to specific principles that ensure its ethical and respon- sible application. These include professional independence, guaranteeing autonomous methodological de- cisions; impartiality and objectivity, preventing discriminatory biases; statistical confidentiality, protecting data privacy; and reliability, ensuring the accuracy and consistency of results. Additionally, transparency allows for the traceability and explainability of the models used, while specificity and proportionality en- sure that AI-ML is used only for justified statistical purposes.

Furthermore, cost-effectiveness assesses the economic and environmental impact of these technologies. Finally, human autonomy and oversight are emphasized, establishing that AI systems cannot make deci- sions without final validation by a qualified human expert. This ethical framework reinforces the commit- ment of public statistics to quality, reliability, and responsibility in the use of AI-ML in official statistics, ensuring alignment with national and international standards.

Acknowledgments: Not provided

Disclosure of Interests: The authors have no competing interests to declare that are relevant to the content of this article. Author 2 is a member of committee of the Ethicomp Congress.

References

1. Gobierno de España. (2024). Estrategia de Inteligencia Artificial 2024. Recuperado de https://www.lamon- cloa.gob.es/consejodeministros/resumenes/Documents/2024/140524-%20Estrategia%20IA%202024.pdf
2. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000379920

OECD. (2019). OECD AI Principles. Retrieved from https://www.oecd.org/going- digital/ai/principles/

3. European Commission. (2019). Ethics Guidelines for Trustworthy AI. Retrieved from https://digital-strategy.ec.eu- ropa.eu/en/library/ethics-guidelines-trustworthy-ai
4. European Union. (2024). Regulation 223/2009 on European Statistics. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02009R0223-20241226
5. European Statistical System. (2022). European Statistics Code of Practice. Retrieved from https://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice

ON THE CURRENT (IM)POSSIBILITY TO ACHIEVE PUBLIC VALUE THROUGH THE EU DIGITAL STRATEGY: AN ETHICS METHOD TO SEEK A "COLLECTUAL" BALANCE

Stefano Calzati[1] [0000-0002-4590-6709]

Keywords: Public Value, Data Ethics, Education.

Extended Abstract

This paper has two aims, one theoretical and one practical: 1) to highlight the criticalities and the ultimate impossibility to achieve public value (singular) by/through digital technologies, based on the current regulatory framework of the European Union; 2) to redress such criticalities by advancing a complementary transdisciplinary approach, realized through a problem-opening method that has been tested in an educational setting at Delft University of Technology.

We repeatedly hear that the digital transformation, especially within/through the public sector, is meant to benefit the collectivity at large. For instance, the release of data as open government data entails the creation of public value – from both governmental organizations and citizens – in the forms (plural) of transparency, accountability, and collaboration [1, 2]. In this respect, various documents part of the EU's digital strategy [3-9] declare the intent to pursue a "people-centric" approach to the digital transformation. But what does this mean? Although quite fuzzy in its conceptualization, "people-centric" entails an exercise of balance between, on the one hand, fundamental human rights – enshrined in the European Charter of Fundamental Rights – and, on the other hand, collective values, such as social inclusion, democratic participation, and environmental sustainability.

The tension between individual and collective standpoints is reflected in the Declaration on Digital Rights and Principles [3], which defends "a European way for the digital transition, putting people at the center." Specifically, the Declaration pins 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. Noteworthy is that these principles equally split between a half (1, 3, 5) focusing on the individual and the other half (2, 4, 6) pertaining to society as a whole. Hence, the Declaration does identify the need to strike a balance between individual and collective standpoints.

However, literature [10-12] has pointed out that what is being enacted in the

1Delft University of Technology, Delft 2628 BX, Netherlands.

EU is, de facto, a human right-based only approach that, while protecting individual freedoms, autonomy, and privacy, tends to overlook the collective-level dimension and effects of the digital transformation, especially its potential societal harms. Hence, the need for striking a "collectual" (collective+individual) balance in the fostering of the digital transformation remains currently unmet from a regulatory normative point of view. To tackle this, it is necessary to complement the current EU's approach with one that a) recognizes the systemic sociotechnical complexity of the scenario we are enmeshed, based on the idea that public value is non-normative strictu sensu [13]; and b) leads to the development of critical competences to be embedded in iterative processes of assessment of such a scenario, if we are to govern it fairly [cf. 14].

On the one hand, theoretical work on public value has shown that such a concept moves well beyond the public sector [13]. This means that public value cannot be reduced to an institutional and legal understanding only, but rather demands the recognition of the full roundedness and contradictory nature of human experience [15] as a relational intersection between subjective evaluation and the societal collective [13]. In connection to that, work in behavioral ethics has highlighted the multifaceted nature of people's decision-making processes whenever they are confronted with moral dilemmas [16]. 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 [17]. Hence, a regulatory normative focus on the individual in relation to the digital transformation is insufficient and must be accompanied with a real-world exploration of how people use digital technologies, why, and with which consequences at collective level.

On the other hand, as works in philosophy of technology have made clear, technology is a pharmakon [18, 19] – both poison and antidote – or, as Kranzberg's [20] first law puts it, "Technology is neither good nor bad, nor is it neutral". At all times, technology can have both positive and negative (un)intended consequences whenever it is applied in context, insofar as its uses generate value-laden entanglements that cannot be taken apart. This implies, in other words, a fractal understanding of public value as irreducible to essential categorizations and demanding, instead, ongoing negotiation and hierarchization of its various (collectual) facets and downsides. There cannot be, for instance, openness (e.g., of data) and transparency, without acknowledging and tackling inevitable closures and opacities. Similarly, a digital service designed to promote inclusive participation might achieve that for certain people and not for others; or it might be inclusive for certain people under certain conditions, but then result exclusive for these same people in later occurrences.

To operationalize a fractal understanding of public value, what is needed is a transdisciplinary approach able to cut across epistemological boundaries, resists any privilege point of reference, and configures an ongoing multidimensional analysis – or, as I call it, an uncertainty-based "problem-opening" method. Such a method is meant to foster an open-ended explorative exercise about the non-neutral impact of digital technologies in any given context. As such, this exercise can be regarded as complementary to the regulatory normative framework of the EU. The paper provides an example of this exercise, by discussing its application and results in a master course on "Data Ethics

for the City" taught over the last three years at TU Delft. Notably, the course was based on three pillars – a sociotechnical understanding of data, the city as a complex system, and ethics as a non-normative non-axiomatic practice – and led students to 1) identify a case study about the implementation and use of a given digital technology in an urban setting; 2) unpack the value-laden entanglements and un/intended consequences of such implementation and use, through literature review and data collection; 3) expose and/or redress such entanglements and consequences through the design of a digital or physical artefact (to be exhibited).

References

1. Twizeyimana, J. D., Andersson, A. The public value of E-Government: a literature review. Government Information Quarterly 36(2), 167-178 (2019)
2. 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. Government Information Quarterly 41(2), https://doi.org/10.1016/j.giq.2024.101930 (2024)
3. European Commission. Declaration of digital rights and principles. https://digital-strategy.ec.europa.eu/en/library/declaration-european-digital-rights-and-principles (2022)
4. European Parliament and Council. European charter of fundamental rights. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT (2012)
5. European Parliament and Council. General data protection regulation. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504&from=EN (2016)
6. European Parliament and Council. A European strategy for data. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066 (2020a)
7. European Parliament and Council. Data governance act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868 (2020b)
8. European Parliament and Council. Data act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN (2022)
9. European Parliament and Council. AI act. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf (2024)
10. Viljoen, S. A relational theory of data governance. Yale Law Journal 131, 573 (2021)
11. Smuha, N. A. Beyond the individual: governing AI's societal harm. Internet Policy Review 10(3) (2021)
12. Calzati, S., van Loenen, B. Beyond federated data: a data commoning proposition for the EU's citizen-centric digital strategy. AI & Society 21, 1-13 (2023).
13. Meynhardt, T. Public value inside: What is public value creation? International Journal of Public Administration 32(3-4), 192-219 (2009)
14. Rochel, J. Ethics in the GDPR: a blueprint for applied legal theory. International Data Privacy Law 11(2), 209–223 (2021)
15. Talbot, C. Paradoxes and prospects of 'public value'. Public Money & Management 31(1), 27-34 (2011)
16. Prentice, P. Teaching behavioral ethics. Journal of Legal Studies Education 31(2), 325–365 (2014)
17. Pink, S., et al. Automation in electric vehicle futures. Mobilities 12, 1–19 (2024)
18. Derrida, J. Dissemination. Athlone Press, London (1981)
19. Lemmens, P. An interview with Bernard Stiegler. Krisis 1, 33–41 (2011)
20. Kranzberg, M. Technology and history: Kranzberg's laws. Technology and Culture 27(3), 544–560 (1986)
21. Calzati, S., Ploeger, H. Problem-solving? No, problem-opening! A method to reframe

Introduction

ADOPTING MORAL SENSITIVITY FOR EVALUATION OF TRUSTWORTHY AI FRAMEWORKS

Adrian Gavornik[1] [0000-0001-5678-0083], Juraj Podrouzek[2] [0000-0002-9691-0310]

Keywords: Trustworthy AI · Moral sensitivity · Ethical frameworks · Responsible AI.

Trustworthy AI and Moral Sensitivity

Recent developments in artificial intelligence have caused growing recognition of the need to develop robust ethical frameworks and governance mechanisms that aim to ensure that the risks related to the development, deployment and use of these technologies is limited. This growing recognition is evident across various sectors, from governmental initiatives to corporate efforts and international organizations. In response, numerous sets of ethical guidelines and principles for trustworthy AI have been developed by governments, companies, and academic institutions.

Meta-analyses of guidelines for trustworthy AI reveal overlap around core principles such as transparency, justice and fairness, non-maleficence, responsibility, and privacy [3–5,7]. Still, it remained a challenge to implement these high level principles into practices, resulting in the so-called "principle-to-practice gap" in AI ethics. In response to the identified gap, several tools and frameworks have been developed to operationalize AI ethics principles. For instance, the Assessment List for Trustworthy Artificial Intelligence provides a checklist for selfassessment, aiming to guide developers in implementing ethical principles [2]

Despite these efforts, questions remain about the actual usage, target audience, and impact of these tools. [1] note that many frameworks lack clarity regarding their intended users and the specific contexts in which they should be applied. Moreover, there is limited empirical evidence assessing the effectiveness of these tools in fostering ethical AI practices. Analyses of ethics toolkits and frameworks suggest limited practical impact [8]. Even widely adopted assessment tools like algorithmic impact assessments have been criticized for potentially becoming bureaucratic exercises that fail to meaningfully influence development practices [6]. The problem extends beyond the tools themselves to questions of measurement and evaluation. While considerable effort has gone into developing operational frameworks, little attention has been paid to assessing their actual use and effectiveness. As [16] note, there is often lack of empirical evidence about whether existing approaches actually improve ethical outcomes or enhance practitioners' capacity for ethical decision-making.

1 Kempelen Institute of Intelligent Technologies, Sky Park Offices, Bottova 7939/2A, 811 09 Bratislava, Slovakia adrian.gavornik@kinit.sk
2 Kempelen Institute of Intelligent Technologies, Sky Park Offices, Bottova 7939/2A, 811 09 Bratislava

If we want to avoid falling into the trap that AI ethics becomes useless [9] or toothless [12], we should find meaningful and sustainable ways to assess the effects these tools have on their users, in this case the AI practitioners. To contribute to a more grounded and empirical assessment of the actual use and effect of existing AI ethics frameworks and guidelines we propose to adopt the concept of moral sensitivity for evaluation of the effectiveness of existing ethical frameworks and tools. Rather than focusing solely on adoption or compliance, this approach examines how well tools and practices develop practitioners' capacity to recognize, evaluate, and address ethical issues in their work. In particular, it examines whether tools and practices enhance practitioners' ability to: (1) identify moral problems; (2) evaluate the relative importance of different ethical concerns; (3) frame moral problems within a broader landscape of ethical principles and values; (4) consider diverse stakeholder perspectives; (5) navigate principle or value conflicts and trade-offs; (6) understand potential moral consequences of actions; and (7) make contextually appropriate decisions.

This approach also takes into account the complex, collaborative nature of AI development, where moral considerations often emerge from team interactions rather than individual decisions. It also provides a way to assess whether tools support what [14] term "situated" ethical decision-making - the ability to recognize and address ethical issues within specific development contexts. The adoption of moral sensitivity has particular relevance given recent findings about the limitations of current approaches which suggest that existing ethics tools often fail to account for the organizational and social contexts in which they're used [11] and that practitioners often struggle to connect abstract ethical principles with concrete development decisions [15]. While moral sensitivity is often discussed at the individual level, ethical decision-making in AI development is significantly shaped by institutional structures and incentives. Thus, it should be taken into account that even when practitioners possess strong moral sensitivity, they may struggle to act on ethical concerns if organizational cultures prioritize efficiency, profitability, or technological innovation over ethical deliberation [10] or feel constrained and lack authority to act [7]. There might even be competing perspectives and priorities across multiple teams and functions within one organization [13]

We aim to further the research agenda with two complementary approaches (Table 1). First, we conduct a scoping survey to understand stakeholder's usage of existing AI ethics tools and frameworks. Building onto these insights, we propose further empirical and qualitative investigations into how the frameworks and tools help foster moral sensitivity in AI practitioners in various contexts, while also paying attention to potential major blockers that prevent them from using and implementing these tools more effectively. Second, we build onto the insights gathered from the scoping survey on the most widespread and used AI ethics frameworks in order to analyse them through the lens of moral sensitivity.

Adopting Moral Sensitivity For Evaluation of Trustworthy AI Frameworks

We want to ask to what extent are the individual components of moral sensitivity addressed in these existing methodologies. It should be noted that even if the strategies are articulated in these frameworks, we might still lack the understanding of their

actual effects in practice. This will provide insight into both what their actual effects on AI practitioners could be but also create space for further improvements, so that trustworthy AI frameworks are not mere checkbox exercises but rather create and foster meaningful changes in AI practice.

Table 1. A table showing the complementary approaches proposed in our future research directions.

Approach	Focus	Methodology	Expected Outcome
Top-down Analysis of AI Ethics Tools and Frameworks	Examines whether and how existing AI ethics guidelines and tools address components of moral sensitivity.	Systematic review of ethical frameworks, tools and guidelines, assessing the occurrence of strategies that could foster specific moral sensitivity components.	Identifies gaps in existing ethics tools and provides insights into their theoretical capacity to foster specific components of moral sensitivity.
Bottom-up Study of AI Practitioners' Moral Sensitivity	Investigates how moral sensitivity manifests in practice and how AI practitioners engage with ethical frameworks and tools.	Qualitative studies (interviews, workshops, ethics interventions) to assess practitioners' moral sensitivity in real-world settings.	Reveals whether selected AI ethics frameworks and tools actually influence practitioners' ethical awareness and behavior, highlighting potential mismatches between intended and actual effects.

Acknowledgments: This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-22-0323 (Philosophical and methodological challenges of intelligent technologies) and ALFIE, a project funded by Horizon Europe research and innovation programme under GA No. 101177912.

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

References

1. Ayling, J., Chapman, A.: Putting ai ethics to work: Are the tools fit for purpose?AI and Ethics (2021). https://doi.org/10.1007/s43681-021-00084-x
2. European Commission. Directorate General for Communications Networks, Content and Technology: The assessment list for trustworthy artificial intelligence (altai) for self assessment. Tech. rep., Publications Office of the European Union (2020). https://doi.org/10.2759/791819
3. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M.: Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. Tech. rep., Social Science Research Network (2020). https://doi.org/10.2139/ssrn.3518482, sSRN Scholarly Paper ID 3518482
4. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V.,Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al.: Ai4people - an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. Tech. rep., SSRN (2018), https://papers.ssrn.com/abstract=3284141, sSRN Scholarly Paper 3284141

5. Jobin, A., Ienca, M., Vayena, E.: The global landscape of ai ethics guidelines.Nature Machine Intelligence 1(9), 389–399 (2019). https://doi.org/10.1038/s42256019-0088-2
6. Metcalf, J., Moss, E., Watkins, E.A., Singh, R., Elish, M.C.: Algorithmic impactassessments and accountability: The co-construction of impacts. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 735–746 (2021). https://doi.org/10.1145/3442188.3445935
7. Mittelstadt, B.: Principles alone cannot guarantee ethical ai. Nature Machine Intelligence 1(11), 501–507 (2019). https://doi.org/10.1038/s42256-019-0114-4
8. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics 26(4), 2141–2168 (2020). https://doi.org/10.1007/s11948-019-00165-5
9. Munn, L.: The uselessness of ai ethics. AI and Ethics (2022). https://doi.org/10.1007/s43681-022-00209-w
10. Orr, W., Davis, J.L.: Attributions of ethical responsibility by artificial intelligence practitioners. Information, Communication & Society 23(5), 719–735 (2020). https://doi.org/10.1080/1369118X.2020.1713842
11. Rakova, B., Yang, J., Cramer, H., Chowdhury, R.: Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. Proceedings of the ACM on Human-Computer Interaction 5(CSCW1), 1–23 (2021). https://doi.org/10.1145/3449081
12. Rességuier, A., Rodrigues, R.: Ai ethics should not remain toothless! a call tobring back the teeth of ethics. Big Data & Society 7(2), 205395172094254 (2020). https://doi.org/10.1177/2053951720942541
13. Schiff, D.S., Kelley, S., Camacho Ibáñez, J.: The emergence of artificial intelligence ethics auditing. Big Data & Society 11(4), 20539517241299732 (2024). https://doi.org/10.1177/20539517241299732
14. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 59–68 (2019). https://doi.org/10.1145/3287560.3287598
Adopting Moral Sensitivity For Evaluation of Trustworthy AI Frameworks 5
15. Vakkuri, V., Kemell, K.K., Jantunen, M., Halme, E., Abrahamsson, P.: Eccola —a method for implementing ethically aligned ai systems. Journal of Systems and Software 182, 111067 (2021). https://doi.org/10.1016/j.jss.2021.111067
16. Yeung, K., Howes, A., Pogrebna, G.: Ai governance by human rights-centred design,deliberation and oversight: An end to ethics washing. Tech. rep., Social Science Research Network (2019). https://doi.org/10.2139/ssrn.3435011, sSRN ScholarlyPaper 3435011

TOWARDS A VALUE SENSITIVE DESIGN (VSD) BASED LEARNING TOOL TO FACILITATE ETHICAL CONSIDERATIONS WHEN DEVELOPING RESPONSIBLE TECHNOLOGIES

Anisha T J Fernando[1] [0000-0001-8974-6906], Kathy Darzanos[2] [0000-0002-3457-1878], Nina Evans[3] [0000-0002-6028-0142], Siaw Mei Sim[4] [0000-0002-7432-6260]

Keywords: Value Sensitive Design, Learning Tools, Ethical Considerations, Value Tensions, Responsible Technology.

Extended Abstract

The world has changed over the past decades due to rapid advancements in technology, globalisation, and shifts in social, cultural, and environmental landscapes. These changes have led to a renewed focus on ethics and human values in all human endeavours. With the increased reliance on technology comes a heightened awareness of ethical challenges that arise, such as privacy concerns, digital surveillance, misinformation, and the potential for deepfakes and data breaches. As diverse cultures regularly interact in globalised environments, it is crucial to promote mutual respect, inclusivity, and understanding.

Responsible innovation incorporates ethical considerations into the development of new technologies, products, services, and systems (Grunwald 2014). Such innovation ensures that ethics is applied practically and is relevant to the innovation context and encourages stakeholder participation when developing technologies (van den Hoven 2014). Responsible technologies are necessary to ensure that technological progress benefits everyone, reduces harm, and is aligned with societal values. Such technologies refer to the development and application of technology in ways that prioritise ethical considerations, sustainability, social impact, and long-term consequences.

An example of unethical conduct when relying on technologies is the Robodebt automated debt recovery program implemented by the Australian government between 2016 and 2019 (Commonwealth of Australia 2023). The scheme aimed to identify and recover overpayments made to welfare recipients by comparing reported income to tax office data. The program used a flawed algorithm leading to incorrect debt calculations, especially for people with variable incomes across a financial year. People were issued

1 UniSA STEM, University of South Australia, Mawson Lakes Campus, South Australia 5095, Australia. anisha.fernando@unisa.edu.au
2 UniSA STEM, University of South Australia, Mawson Lakes Campus, South Australia 5095, Australia. kathy.darzanos@unisa.edu.au
3 UniSA STEM, University of South Australia, Mawson Lakes Campus, South Australia 5095, Australia. nina.evans@unisa.edu.au
4 UniSA STEM, University of South Australia, Mawson Lakes Campus, South Australia 5095, Australia. siaw.sim@unisa.edu.au

automated debt notices without human review, leading to stress and financial hardship for the affected individuals. A class action was filed, and the government agreed to settle for $1.2 billion. The Robodebt scandal highlights the risks and ethical challenges related to automated decision-making processes (Clarke, Michael & Abbas, 2024).

Incorporating ethical considerations in the design and development of technology is crucial to ensure that responsible technologies contribute positively to society without causing harm. Such ethical considerations include, amongst others:

- Fairness: Provide equal access and opportunities for all, regardless of background or identity.
- Well-being: Prioritise the mental, emotional, and physical well-being of users.
- Autonomy: Give users control over how they interact with technology and how their data is used.
- Security: Minimise risks like cybersecurity threats, misinformation, and abuse.

Value Sensitive Design (VSD) is predicated on the principle that technology should not only meet functional needs but also reflect and respect the values of individuals and communities impacted by the technology. VSD therefore considers the core values to uphold (such as fairness and equity, privacy and security, transparency, as well as sustainability) when designing and using technology. VSD offers opportunities to consider ethical implications such as through moral imagination techniques when developing responsible innovations (Umbrello 2020). Technology design influences user behaviour and enables users to practice ethics through the human values embedded within the design (Verbeek 2011; Vallor 2018). VSD considers values that may be in conflict in the design of technology, referred to as 'value tensions'. These value tensions lead to ethical dilemmas, where choosing one value may compromise another (Friedman & Hendry 2019). Recognising and addressing these tensions can lead to improved ethical considerations in technology design.

Learning tools that consider values enable participants to partake in ethical decision-making, which is particularly important when creating responsible technologies. The harm caused by unethical conduct involving technologies can be reduced when individuals are empowered to voice value tensions. For this reason, it is important to empower students to observe and voice value tensions as they become more professional in their conduct. Learning tools that enable students to observe, discuss and knowingly apply values relevant across professions are scarce. Therefore, there is a need for teaching and learning resources that afford opportunities to identify, consider and discuss values that are relevant to the appropriate tertiary education context.

In this study, we conducted a scoping review and used a mixed methods action research strategy using the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) approach. The scoping review explored VSD-based learning tools that embed values and consider ethical perspectives during the design and development of responsible technologies. We described ethical considerations that can be applied when developing responsible technologies and explore how they can be embedded in learning tools.

We investigated the use of a learning tool in the form of VSD conversation cards to explore the value tensions between social and market-based norms at play using technology (Fernando, 2020; Fernando & Scholl 2020). The VSD conversation cards were originally developed to enable information technology (IT) students to observe, reflect, and discuss value tensions in the Australian tertiary education context. These cards were refined to extend value tensions by embedding Vallor's (2018) techno-moral virtues (social values) and Zuboff's (2019) analysis of systemic drivers of innovations (market values). The VSD conversation cards were piloted as part of a larger repeated measures empirical study. The pilot study was conducted in three phases: pre-intervention measurement, an intervention using the conversation cards, and post-intervention measurement. The participants were undergraduate and postgraduate students from a wide range of disciplines. A factorial vignette survey to measure value competencies was designed.

The pre- and post-intervention measurements used a survey to gauge the values competencies of participants. Following the pre-intervention survey, respondents undertook a group-based learning activity (intervention) in tutorials specific to their course to embed a values-based ethics lens to their specific discipline. Students examined scenarios to articulate and explored the tensions in the context of their learning activity. Respondents then completed the survey again, thereby providing a post-treatment measure.

An analysis of variance (ANOVA) is used to gauge the efficacy of the conversation cards to develop values competencies and ethical decision-making skills, and to identify factors carrying the most weight. Selected representative students participate in semi-scripted, videoed focus groups. Thematic analysis is used to elicit key themes and insights from the qualitative focus group data.

After the pilot study, further empirical research will be conducted across several courses and disciplines to validate the conversation cards. This research will enable the development of teaching and learning resources (the learning tool) for different disciplines. We also aim to propose an innovative teaching and learning approach to develop students' ethics competencies by observing and discussing value tensions through VSD. This approach aims to engage multi-disciplinary educators to co-design learning activities with ethical dilemmas related to the design and development of IT within their professions using the VSD conversation cards. These co-designed learning activities and the VSD conversation cards will be made available to educators via a shared repository. The study aims to provide future students with greater access to diverse multi-disciplinary scenarios for technology design, development, and use. An online version of the validated conversational cards will also be developed to extend accessibility to online students via an open educational resource (OER).

Responsible innovation relates to the appropriate application of ethical considerations when developing technologies. Value sensitive design offers an important avenue for students to learn about, consider and discuss ethical dilemmas that may arise in creating these technologies. The study uses mixed methods action research to conduct a scoping review and a repeated measures empirical study to identify important ethical considerations that need to be included in the development of learning tools. A key proposed outcome of this study is to develop an online version (OER) of the validated VSD

conversation cards by incorporating Vallor's (2018) techno-moral virtues (social values) and Zuboff's (2019) analysis of systemic drivers of innovation (market values).

Acknowledgments: This research is supported by a University of South Australia Unstoppable Teaching and Learning Development Grant (2025-2026).

Disclosure of Interests: The authors have no competing interests.

References

1. Clarke, R., Michael, K. & Abbas, R.: Robodebt: A Socio-Technical Case Study of Public Sector Information Systems Failure. Australasian Journal of Information Systems, 28, 1-42 (2024)
2. Commonwealth of Australia, Royal Commission into the Robodebt Scheme. Final Report vol 1-3 (2023)
3. Fernando, A.: Exploring How Value Tensions Could Develop Data Ethics Literacy Skills. In: Pelegrin-Borondo, J., Arias-Oliva, M., Murata, K., Lara Palma, A.M. (eds.), ETHICOMP 2020 CONFERENCE: 18th International Conference on the Ethical and Social Impacts of ICT (pp. 194 - 197). Universidad de La Rioja (2020)
4. Fernando, A.T. & Scholl, L.: Towards using value tensions to reframe the value of data beyond market-based online social norms, Australasian Journal of Information Systems, 24 (2793), 1-9 (2020)
5. Friedman, B., & Hendry, D.G.: Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press, Cambridge, MA (2019)
1. Grunwald, A.: Technology Assessment for Responsible Innovation. In: van den Hoven, J., Doorn, N., Swierstra, T., Koops, BJ., Romijn, H. (eds.) Responsible Innovation 1. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-8956-1_2 (2014)
2. Umbrello, S.: Imaginative Value Sensitive Design: Using Moral Imagination Theory to Inform Responsible Technology Design. Science and Engineering Ethics, 26(2), 575–595. https://doi.org/10.1007/s11948-019-00104-4 (2020)
3. Vallor, S.: Technology and the virtues: A philosophical guide to a future worth wanting. Oxford University Press, New York (2018)
4. van den Hoven, J.: Responsible Innovation: A New Look at Technology and Ethics. In: van den Hoven, J., Doorn, N., Swierstra, T., Koops, BJ., Romijn, H. (eds.) Responsible Innovation 1. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-8956-1_1 (2014)
5. Verbeek, P.-P.: Moralizing technology: understanding and designing the morality of things, University of Chicago Press, Chicago (2011)
6. Zuboff, S.: The age of surveillance capitalism: the fight for the future at the new frontier of power. Profile Books, London (2019)

ARTIFICIAL MORAL DISCOURSE AND THE FUTURE OF HUMAN MORALITY: MECHANISMS OF INFLUENCE

ELIZABETH O'NEILL[1][0000-0002-2794-498X]

Keywords: moral machines, artificial moral advisors, LLM-based chatbots, moral psychology, AI ethics, artificial moral discourse

Extended Abstract

The moral norms, values, and practices of *Homo sapiens* have not always been as they are now. There was a causal process by which humans came to care about values like peace, generosity, privacy, and justice, as well as values like honor, chastity, piousness, and purity, and, more generally, the question of what one ought to do. Likewise, there was a causal process by which we developed our various morality-related practices—things like justification, scolding, mercy, promising, and so on. These aspects of human moral life are not set in stone—they may change in the future. (For general work on the topic of moral change, see FitzPatrick 2021, Sterelny 2021, Baker 2019, Tomasello 2016). We may acquire new values and norms and we may lose or come to care less about some of those that we currently have. Similarly, it's possible that the moral practices that humans currently engage in will look dramatically different in the future.

One striking new way in which human morality might change is through human interactions with AI systems. In particular, we can think of interactions with chatbots like ChatGPT—conversational systems, based on large language models (LLMs), that are easily accessible to the public and growing in use—and, more broadly, multimodal AI systems based on foundation models.

Many publicly-accessible large language model (LLM)-based chatbots readily and flexibly generate outputs resembling moral assertions, advice, praise, expression of moral emotions, and other morally-significant communications. We can call this phenomenon "artificial moral discourse" (O'Neill and Nyholm manuscript). In the first part of this talk, I offer a characterization of artificial moral discourse, according to which:

> To engage in artificial moral discourse is for a computer system to:
> exhibit a pattern of response to inputs that resembles some human pattern of response to similar inputs, where the response contains something (e.g., terms, sentences, gestures, facial expressions, etc.) that the human interlocutor (or an observer) views as communicating a moral message or would have viewed as communicating a moral message if the exchange had occurred between humans (O'Neill and Nyholm manuscript).

Drawing from existing empirical studies (e.g., Howe et al. 2023, Fisher et al.

1 Philosophy & Ethics, Eindhoven University of Technology, Eindhoven, The Netherlands, e.r.h.oneill@tue.nl

preprint), I provide several examples of artificial moral discourse.

Why should we be interested in artificial moral discourse? Among other things, interactions with artificial moral discourse could influence human moral values, norms, character traits, and practices, for good or ill (O'Neill and Nyholm manuscript). In the second part of the talk, I make a case for the claim that artificial moral discourse is likely to influence human morality in ways that past technologies have not. Namely, I propose that regular interaction with LLM-based chatbots can influence human moral cognition and practices via mechanisms that resemble modes of social influence on morality. This includes influence via advice and testimony, behavioral influence, and influence via norm enforcement.

The first mechanism I discuss is chatbot influence resembling moral advice and testimony (Wiland 2021, Krügel et al. 2023). Humans may seek out this advice and testimony, or they may simply encounter it in the course of their exchanges with the system. Outputs resembling testimony could influence human norms and values via deference, in which the human accepts the statement of the chatbot on its apparent say-so, or by altering human perceptions of what typical views are, inclining them to alter their own views to be more in alignment. Influence could also come about via humans' oppositional reactions to the systems' outputs: especially when humans perceive the chatbots' view as imposed on them from above or as a view espoused by a source they distrust, they may modify their own view to be more opposed to the view they encounter from the chatbot. Outputs resembling moral advice may influence human morality in a slightly more indirect way. Humans may simply accept the moral advice of the chatbot or they may act in a contrarian way; either way, the decisions they make in turn will influence the values they hold and will influence others' perceptions of what norms prevail and what actions are acceptable.

The second mechanism I discuss is behavioral influence, including influence via example and imitation (Henrich 2016, Lockwood et al. 2025, Köbis et al. 2021). The idea here is that via mechanisms analogous to the ways in which humans are influenced by other humans' mannerisms, word choices, tone of voice, ways of communicating, and other behaviors, humans that interact with LLM-based chatbots may also modify their behaviors so as to be more like the AI system, or contrarily, to be *unlike* the AI system. Already, it is easy to imagine people modifying how they communicate based on what they've observed chatbots doing during their text or voice exchanges. Speculatively, if video-generation capacities improve enough that humans regularly interact with chatbots with video personas, we could also imagine people imitating the facial expressions and visually observable mannerisms that the systems use. This might not immediately seem morally significant, but humans often use such indicators to convey morally significant messages, such as, e.g., concern, lack of caring, cruel sarcasm, or the message that something's just a little off in what another person just said or did. These are all things where we can imagine that if humans routinely interact with the chatbots, humans' own dispositions for what to do and say and how to react might change.

The third mechanism I discuss is automated norm enforcement. In sum, there are some properties and actions that some humans consider to be wrong, which humans are liable to think that computers could detect on the basis of imagery or sound, given the ability of the systems to imitate patterns of human moral language usage. For instance,

in 2023, Iranian officials announced they would use AI in public spaces to detect women whose hair is uncovered. Officials could subsequently target those women for fines, arrest, or "morality schooling" (Johnson 2023). This news item is a dramatic instance of a larger trend. With advancements in computer vision systems have come efforts to adapt those systems to identify a variety of norm violations, including violence and aggression (Sernani et al. 2021), loitering (Shanmughapriya et al. 2022), littering (Brandon 2022), use of force by police officers (Farooq 2024), or simply "unusual behavior" (Fickenscher 2023). Given the important role that sanction and reward play in human normative life, the use of AI systems to identify norm violation and compliance—whether to feed into human decisions about sanction and reward or to automatically mete out such responses—is highly significant. If AI systems are harnessed in this way, we can imagine people changing their behaviors to evade the system's artificial gaze, which may correctly or incorrectly classify them as engaging in a particular immoral action. Changing behavior can lead to changing norms and eventually values.

In some ways, artificial moral discourse resembles the ideas of moral machines and artificial moral advisors, which inspired a lively literature over the last few decades (Wallach and Allen 2008, Anderson and Anderson 2011). Despite this, the theoretical resources developed for thinking about those hypothetical phenomena are insufficient for guiding our analysis and evaluation of artificial moral discourse. With LLM-based chatbots, it is far from clear that the systems in question are engaging in genuine moral reasoning, advice-giving, etc., nor can we assume that they are *reliable* sources of moral judgments or advice. Indeed, given the systems' complexity, opacity, and novelty, we can say little with confidence about their behavioral dispositions or about the conditions that elicit particular behaviors from them. In fact, there currently exist no well-validated tests or standards for evaluating their morally-relevant capacities across a range of contexts. In the third part of the talk, I suggest some further research questions for future empirical, technical, and philosophical investigation on how artificial moral discourse may influence human morality and what the ethical implications of that influence may be.

Acknowledgments. This work was supported by the Netherlands Organization for Scientific Research (NWO) grant #406.XS.03.070, "When computers join the moral conversation." This work was also performed in the context of the research programme Ethics of Socially Disruptive Technologies, which is funded through the Gravitation programme of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO Grant Number 024.004.031).

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

References

1. Anderson, M. and S.L. Anderson (eds.) 2011. *Machine ethics*. Cambridge University Press.
2. Baker, R. 2019. *The structure of moral revolutions: Studies of changes in the morality of abortion, death, and the bioethics revolution*. MIT Press.
3. Brandon, E.M. January 12, 2022. England has an £850 million problem with litter. This startup is using AI to fix it. Fast Company. https://www.fastcompany.com/90711804/england-has-an-850-million-problem-with-litter-this-startup-is-using-ai-to-fix-it
4. Farooq, U. Feb 2, 2024. Police Departments Are Turning to AI to Sift Through Millions of Hours of Unreviewed Body-Cam Footage. ProPublica. https://www.propublica.org/article/police-body-cameras-video-ai-law-enforcement

5. Fickenscher, L. Feb 12, 2023. Retailers busting thieves with facial-recognition tech used by MSG's James Dolan. New York Post. https://nypost.com/2023/02/12/retailers-busting-thieves-with-facial-recognition-tech-used-at-msg/
6. Johnson, K. Jan 18, 2023. Iran says face recognition will ID women breaking hijab laws, Wired. https://www.wired.com/story/iran-says-face-recognitionwill- id-women-breaking-hijab-laws/.
7. Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D.W., Pan, J., Tsvetkov, Y. and Reinecke, K. Preprint, 2024. Biased AI can influence political decision-making. *arXiv preprint arXiv:2410.06415*.
8. FitzPatrick, W. 2021. Morality and Evolutionary Biology. *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2021/entries/morality-biology/.
9. Henrich, J. 2016. The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter. Princeton University Press.
10. Howe, P.D.L., Fay, N., Saletta, M. and Hovy, E. 2023. ChatGPT's advice is perceived as better than that of professional advice columnists. *Frontiers in Psychology*, *14*, p.1281255.
11. Köbis, N., Bonnefon, J.F. and Rahwan, I. 2021. Bad machines corrupt good morals. *Nature Human Behaviour*, *5*(6), pp.679-685.
12. Krügel, S., Ostermaier, A. and Uhl, M. 2023. ChatGPT's inconsistent moral advice influences users' judgment. Scientific Reports, 13(1), p.4569.
13. Lockwood, P.L., van den Bos, W. and Dreher, J.C. 2025. Moral learning and decision-making across the lifespan. *Annual Review of Psychology*, *76*(1), pp.475-500.
14. O'Neill, E., & Nyholm, S. Manuscript. The Advent of Artificial Moral Discourse: LLM-based AI chatbots mimicking human moral discourse.
15. Sernani, P., Falcionelli, N., Tomassini, S., Contardo, P. and Dragoni, A.F. 2021. Deep learning for automatic violence detection: Tests on the AIRTLab dataset. IEEE Access, 9, pp.160580-160595.
16. Shanmughapriya, M., Gunasundari, S. and Bharathy, S. 2022, April. Loitering detection in home surveillance system. In: 202210th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-22) (pp. 1-6). IEEE.
17. Sterelny, K. 2021. *The Pleistocene social contract: Culture and cooperation in human evolution*. Oxford University Press.
18. Tomasello, M. 2016. *A natural history of human morality*. Harvard University Press.
19. Wallach, W., & Allen, C. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.
20. Wiland, E. 2021. Guided by voices: Moral testimony, advice, and forging a 'we'. Oxford University Press.

STAKES, CONTEXT, AND AI: HOW AI CHALLENGES HUMAN WAYS OF KNOWING

ANDREW P. REBERA[1] [0000-0003-1148-5755]

Keywords: artificial intelligence, evidentialism, moral encroachment, pragmatic encroachment, testimony

Extended Abstract

Artificial Intelligence systems, particularly Large Language Models (LLMs) and generative AI, now play a prominent role in our everyday epistemic practice—serving as sources of knowledge and ideas. While the ethics literature on these technologies remains unsettled, the most commonly cited areas of concern [1]—fairness, bias, safety, harmful content, privacy—all suggest the centrality of epistemic considerations. The epistemology of AI is therefore increasingly urgent and fundamental.

In this paper I examine the potential impact of the increasingly prominent role of generative AI and LLMs (henceforth just "AI" for short) in our epistemic practice. I argue that AI systems function as agents in our epistemic networks, i.e. the systems of knowledge-sharing among agents, operating with an inherently "evidentialist" approach that struggles to accommodate what I call the "pragmatic insight"—the recognition, commonplace in the contemporary epistemology literature, that practical and moral stakes can influence epistemic standards. I suggest that this limitation carries two significant risks: first, that deployment decisions may fail to account for AI's epistemic limitations; and second, more fundamentally, that human epistemic practices may gradually come to conform to AI's limitations, potentially transforming how we understand ourselves and each other.

Stage one of my argument is to clarify AI's role in our epistemic practice. As recently noted [2], beliefs acquired from AI might seem to occupy an in-between position, conforming to neither standard accounts of instrument-based beliefs nor testimony-based beliefs. I suggest, however, that it is right to conceptualize AIs as agents in our epistemic networks and, therefore, to treat at least some of their output as grounding testimony-based beliefs. I present an argument for this view based on the idea that AIs are subject to epistemic norms when they assert. (I briefly defend the view that AIs are capable of assertion from recent objections due to Butlin & Viebahn [3].)

Having established that AIs are able to function as agents in our epistemic networks, participating in testimonial exchanges (i.e. the transferring of knowledge through testimony or assertion), I turn to my central argument. Recent work on the epistemology

1 DEPARTMENT OF BEHAVIOURAL SCIENCES, ROYAL MILITARY ACADEMY, BRUSSELS, BELGIUM; 2 INSTITUTE OF PHILOSOPHY, KU LEUVEN, BELGIUM

of AI has addressed trust and epistemic authority [4], [5]. My focus is different. One of the central debates in contemporary epistemology is between evidentialists and proponents of pragmatic and moral encroachment. Evidentialism is, roughly, the view that the justification of beliefs depends solely on evidence. This view is challenged by proponents of pragmatic and moral encroachment. According to these views, non-evidential factors can play a determining role in whether a subject's belief constitutes knowledge. In supposed cases of pragmatic and moral encroachment, practical or moral factors—notably what is at stake in the context—influence the standards of evidence required for knowledge [6], [7]. However these debates ultimately play out, both sides in it agree that the phenomenon of practical or moral stakes appearing to influence epistemic standards warrants explanation. I call this the "pragmatic insight."

Stage two in my argument is to show that AIs are, in practice, natural evidentialists and that, as such, they struggle to adequately accommodate the pragmatic insight. The appropriate formation of beliefs about sensitive personal, social, ethical or political issues often requires careful attention to context and social dynamics. For example, proponents of moral encroachment argue that people should not rely solely on statistical reasoning when doing so could have serious consequences, such as in cases of racial profiling [7]. Yet AI systems, relying on identification and analysis of statistical patterns in training data, systematically fail to make such adjustments. They process all evidence through the same statistical frameworks, regardless of the stakes involved. For example, in responding to medical queries, LLMs apply the same confidence thresholds regardless of the seriousness of medical conditions they are writing about or the likely impact on the person asking; or again, when prompted to say something about a particular social group, an LLM will generate responses based on statistical patterns without recognizing when its output might constitute harmful stereotyping—exactly the kind of case where humans (one hopes) would act more sensitively, observing higher evidential standards. Hard-coding can address some such cases, but of course hard-coding is by nature context-insensitive.

Measures to address this limitation could be attempted through either "top-down" approaches (e.g. explicit rules about adjusting confidence thresholds based on stakes) or "bottom-up" approaches (e.g. training on data that reflects human context-sensitive judgment[2]). However, neither approach is fully plausible. Top-down approaches are problematic because the relevant contextual factors resist codification, while bottom-up approaches may struggle because, e.g., training data cannot fully capture the nuanced ways in which humans adjust their belief-forming processes to contextual factors. Likewise, the proposal that AIs could be informed of the stakes and relevant contextual factors through prompting on a case-by-case basis falls short not only for being wildly impractical, but because it requires of the users who do the prompting an unrealistically high degree of contextual knowledge and practical wisdom. A hybrid approach, combining human oversight in high-stakes domains with systems designed to flag context-dependent cases, may be more promising. This would acknowledge AI systems' evidentialist limitations and appeal to human context-sensitivity where essential.

2 By context-sensitive judgement I mean adjusting epistemic standards according to pragmatic, social, and moral factors in a given situation. This involves recognizing when stakes demand higher evidential standards—a capacity requiring both explicit reasoning and implicit sensitivity that remains difficult to implement in AI systems despite ongoing research.

Stage three of the argument is to highlight two key implications of AI's inability to accommodate the pragmatic insight. The first implication—an anyway familiar observation in AI ethics—is that deployment decisions about AIs should be made in full awareness of this limitation and on how it might play out in different contexts of use. The second implication concerns the risk that as AIs assume more and more prominent roles in epistemic practices, so human epistemic behaviors may come to conform to AI's evidentialist limitations. Ways in which technology can mediate action and experience, including moral action and experience [8], are well documented; but here I advert specifically to AI's impact on epistemic practices—potentially its negative impact whereby subtle adaptations to its evidentialist limitations cause what Bisconti calls a "transfer of dysfunctional patterns" from human-AI to human-human interactions [9, p. 499]. This transfer could result in humans carrying AI's disregard for context-sensitivity into their interactions with each other. An analogous phenomenon is AI-Mediated Communication, which can "subvert existing social heuristics" and challenge norms of personal agency [10]. Our understanding of each other (as well as of ourselves) is at stake. It is not fully clear how humans navigate complex, context-sensitive epistemic processes. If our epistemic practices come to be modeled on AI systems, we risk misunderstanding ourselves and each other, as human belief formation comes to subtly mirror AI's approaches, which are arguably more systematic, but less context-sensitive. This could fundamentally alter how we process and share knowledge within communities.

The argument has significant implications for current debates about AI ethics, epistemology, and development. While much attention focuses on making AIs more accurate, transparent, or aligned with human values, I suggest we must also attend to how their basic epistemic characteristics might influence—and possibly distort—human epistemic practices. This potential transformation of human epistemic practices represents a profound ethical challenge that demands specific responses: (1) designing AI systems with greater awareness of contextual factors, (2) cautious deployment in domains where context- and stakes-sensitivity is crucial, and (3) making AI users more aware of AI systems' limitations. These considerations should be addressed alongside more commonly discussed concerns about AI safety and fairness.

Acknowledgments: This study was funded by a grant from Belgian Defence, HFM 22-03.

Disclosure of Interests: The author has no competing interests to declare that are relevant to the content of this article.

References

1. T. Hagendorff, 'Mapping the Ethics of Generative AI: A Comprehensive Scoping Review', Minds & Machines, vol. 34, no. 4, p. 39, Sep. 2024, doi: 10.1007/s11023-024-09694-w.
2. O. Freiman, 'Analysis of Beliefs Acquired from a Conversational AI: Instruments-based Beliefs, Testimony-based Beliefs, and Technology-based Beliefs', Episteme, vol. 21, no. 3, pp. 1031–1047, Sep. 2024, doi: 10.1017/epi.2023.12.
3. P. Butlin and E. Viebahn, 'AI Assertion', Ergo, forthcoming, [Online]. Available: https://osf.io/preprints/osf/pfjzu
4. R. Alvarado, 'What kind of trust does AI deserve, if any?', AI Ethics, vol. 3, no. 4, pp. 1169–1183, Nov. 2023, doi: 10.1007/s43681-022-00224-x.

5. A. Ferrario, A. Facchini, and A. Termine, 'Experts or Authorities? The Strange Case of the Presumed Epistemic Superiority of Artificial Intelligence Systems', Minds & Machines, vol. 34, no. 3, p. 30, Jul. 2024, doi: 10.1007/s11023-024-09681-1.
6. J. Fantl and M. McGrath, 'Evidence, Pragmatics, and Justification', The Philosophical Review, vol. 111, no. 1, pp. 67–94, Jan. 2002, doi: 10.1215/00318108-111-1-67.
7. S. Moss, 'Moral Encroachment', Proceedings of the Aristotelian Society, vol. 118, no. 2, pp. 177–205, Jul. 2018, doi: 10.1093/arisoc/aoy007.
8. P.-P. Verbeek, What things do: philosophical reflections on technology, agency, and design, 2. print. University Park, Pa.: Pennsylvania State Univ. Press, 2005. Accessed: Dec. 01, 2022. [Online]. Available: http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&doc_number=013143080&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA
9. P. Bisconti, 'How Robots' Unintentional Metacommunication Affects Human–Robot Interactions. A Systemic Approach', Minds & Machines, vol. 31, no. 4, pp. 487–504, Dec. 2021, doi: 10.1007/s11023-021-09584-5.
10. J. T. Hancock, M. Naaman, and K. Levy, 'AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations', Journal of Computer-Mediated Communication, vol. 25, no. 1, pp. 89–100, Mar. 2020, doi: 10.1093/jcmc/zmz022.

THE NEED TO ADDRESS EMERGING AI CHALLENGES IN RESEARCH PRACTICES THROUGH RESEARCH ETHICS REVIEWS

KONSTANTINA GIOUVANOPOULOU[1], XENIA ZIOUVELOU[2], KOSTAS KARPOUZIS[3], VANGELIS KARKALETSIS[4]

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

Technology has the potential to transform our lives and, undoubtedly, the widespread application of Artificial Intelligence (AI) across different domains poses a number of challenges. Therefore, it is vital to reflect on emerging ethical and social implications related to AI across different domains and contexts by adopting a proactive approach for the design, development, and continuous evaluation of AI systems, that is anticipatory in nature. This will allow these challenges to be reflected upon, identified and addressed at an early stage, leaving ample time for researchers and developers to react. Given the fact that the existing AI regulatory framework (EU AI Act) doesn't apply to AI systems intended for scientific research, it is important to adopt a reflective, ethics-by-design approach strongly integrated in all research phases, start-ing from the research design stage. Similarly, the emerging AI challenges are significantly impacting the process and nature of the research ethics review that aims to ensure ethical and responsible research practices. As such, it is essential to explore these challenges in order to better understand the impact of AI and to increase the effectiveness of the research ethics review process for achieving more ethical AI research outcomes with positive societal impact by design.

Aiming to address these concerns, this study explores the challenges to be addressed through the research ethics review process overseeing and reviewing AI related research, given the potential risks for individuals and society. AI-related challenges may involve new ethical concerns or extensions of existing ones, may concern diverse areas and be either scientific, technical, political, legal and social, and arising risks may relate to the safety of applications, the security and privacy of users. Given that the increasing effectiveness of AI is proportional to its social responsibility, AI scientists should raise awareness about the current risks and undertake integrated research so as to address the social risks of AI [1].Research ethics, covering the whole spectrum of research ac-

1INSTITUTE OF INFORMATICS AND TELECOMMUNICATIONS, NATIONAL CENTRE FOR SCIENTIFIC RESEARCH DEMOKRITOS, GREECE; DEPARTMENT OF COMMUNICATION MEDIA & CULTURE, PANTEION UNIVERSITY OF SOCIAL AND POLITICAL SCIENCES, GREECE, KGIOUVANO@IIT.DEMOKRITOS.GR

2 INSTITUTE OF INFORMATICS AND TELECOMMUNICATIONS, NATIONAL CENTRE FOR SCIENTIFIC RESEARCH DEMOKRITOS, GREECE, XENIAZIOUVELOU@IIT.DEMOKRITOS.GR

3 DEPARTMENT OF COMMUNICATION MEDIA & CULTURE, PANTEION UNIVERSITY OF SOCIAL AND POLITICAL SCIENCES, GREECE, KKARPOU@PANTEION.GR

4 INSTITUTE OF INFORMATICS AND TELECOMMUNICATIONS, NATIONAL CENTRE FOR SCIENTIFIC RESEARCH DEMOKRITOS, GREECE, VANGELIS@IIT.DEMOKRITOS.GR

5 EU-AI ACT (REGULATION (EU) 2024/1689)

tivity — from the researcher, the research methods to be applied, the funding of the research and the Research Ethics Committees (RECs) — aims to promote good and to capture emerging challenges. Research ethics review is the process of examining proposed research protocols so as to ensure that the research will be conducted ethically and responsibly, demonstrating compliance with law and ethical standards, protecting all participants, minimising the potential risks that may arise and thus, avoiding societal harm. A typical research ethics review process involves the submission of the research protocol accompanied by all the relevant and necessary documents (e.g. informed consent forms) [2]. Research Ethics Committees (RECs) are then responsible for conducting ethics reviews, operating either in research institutions, governmental organisations or within the private sector [3], holding the crucial role of ensuring that research is designed and will be conducted following national and international laws and aiming to protect research participants [4]. RECs provide a multidisciplinary approach to review, assure compliance with commonly accepted rules, principles, soft law, codes of conduct and monitor the scientific integrity of a project from the early beginning [5].

Considering RECs key role in research ethics it is essential that they follow effective procedures for reviewing. Regarding the AI research domain, its rapid development and the risks it entails challenge the role of RECs and make the evaluation process even more demanding. RECs should be able to proactively detect AI emerging challenges when reviewing AI research proposals so as to foster responsible innovation and promote social good. Examining further the challenges associated with AI technology and in an effort to explore how these challenges are addressed through research ethics reviews, we conducted a systematic literature review addressing the following key research questions:

- 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?

Towards this aim, a systematic literature review was carried out following the systematic literature review guidelines (SLR) by Kitchenham and Charters [6] and PRISMA[6] 2020 guidelines. The research conduct strategy was defined from the outset to detect the relevant literature by specifying the research questions to be addressed, as well as the method to perform the literature review. Scopus database was chosen as a source of relevant articles responding to our research needs and 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 resulted in 169 articles, and excluding any other language than English, 162 articles were finally included for consideration. On 19 August a title detection was carried

6 Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 guidelines, https://www.prisma-statement.org/prisma-2020, last accessed: 28 March 2025.

out and of the 162 articles resulting from the search, 66 were selected using specific inclusion and exclusion criteria. Subsequently, an abstract detection was carried out on 20 August which resulted in 56 articles out of 66 using inclusion and exclusion criteria as well. The full text detection was conducted from 21 August to 1 November 2024 and resulted to 30 documents applying again certain inclusion and exclusion criteria. The articles resulting from the systematic literature review provide important insights into the challenges faced by RECs in general and those posed by AI technology across different sectors. In addition to the systematic literature review, a few other articles that were identified from the literature and are considered particularly influential and instrumental to a more thorough study of the topic were also examined [7]. These selected articles provided input for a more in-depth analysis of certain aspects of the topic under study.

The study of the topic identified diverse levels of examination that enabled us to create a Classification Framework (Table 1) including the following types of challenges:

1. Research Ethics Review Challenges: comprising generic challenges (1a) in relation to research ethics reviews (including structural challenges regarding the composition and the expertise of RECs, and procedural challenges focusing on the review process) as well as AI-related challenges (1b), that need to be addressed through the research ethics review process.
2. Domain specific challenges: emerging AI challenges across different domains (2a), arising from the application of AI technology in different fields (e.g., Information and Communication Technology (ICT), Healthcare, Education, Judiciary).

The literature survey carried out reveals that there are many articles that address the ethical implications of the rapid development of AI in diverse fields, while fewer articles deal with the generic challenges faced by RECs conducting ethics reviews and how AI challenges are addressed through the ethics review process. Regarding the European level, RECs differ in structure and the procedures they follow and therefore, different models can be detected per country, depending on the legal and organisational framework [8]. In general, detected challenges are related to the structure of the committees, their composition, the need for diverse expertise, multi-disciplinarity, and training in ethics and emerging technologies [9], and to the procedures they follow to evaluate research projects. Procedural challenges of evaluating research projects/ proposals relate to issues of bureaucracy [2], weak regulatory framework or over-regulation, inadequate financial and human resources, and inconsistencies in decision-making [3], [10], [11]. Challenges related to the development and use of AI concern different sectors, such as Education [12], Healthcare [13], [14], [15], [16], Judiciary [17] and ICT [18]. It is noteworthy that the AI emerging challenges detected in different areas seem to converge in general points (Table 1). Issues of transparency in decision-making, safety, explainability, data protection and arising dilemmas related to privacy concerns [9] were detected as AI challenges in relation to research ethics and research ethics reviews; and were also detected, in general terms, in the examination of articles dealing with AI emerging challenges in different fields.

Table 1. Research Ethics Review and AI Challenges: A Classification Framework

Types of Challenges			Identified Challenges related to:	Indicative Sources
1. Research Ethics Review Challenges	1a. Generic Challenges	Structural Challenges	RECs composition, need for diverse expertise, multidisciplinarity, lack of training in research ethics and/emerging technologies	[9]
		Procedural Challenges	Bureaucratic issues, Weak regulation/Over-regulation, Lack of financial, human resources, Decision-making inconsistencies	[2], [3], [10], [11]
	1b. AI-related Challenges	AI-Challenges for Research Ethics and Research Ethics Review	Transparency, Explainability, Data Protection, Safety, Dilemmas regarding privacy issues vs. scientific progress, Chasm between research ethics and AI research, Obtaining informed consent	[9]
2. Domain Specific Challenges	2a. Emerging AI Challenges across different domains	ICT	Explainability, Data Privacy, Safety, Security, Bias, Fairness	[12], [14], [15], [16], [17], [18]
		Healthcare	Interpretability, Explainability, Gender bias, Accountability, Safety	
		Education	Accessibility, Reliability of AI, Privacy	
		Judiciary	Bias, Fairness, Transparency, Accountability, Privacy, Data Protection	

Research related to AI leads to complex issues and RECs' concerns relate to assessing the ethical issues arising from and adapting to the rapid changes occurring in this emerging field [9]. The implications posed by AI imply new challenges that RECs have to overcome when evaluating AI research. The absence of in-depth research in this area, based on our research findings, creates great scientific interest. It is therefore necessary to explore further this field in order to address how RECs could evaluate AI research proposals more effectively given that they operate under different frameworks at a European and international level. Such research will enable us to reach more ethical AI research outcomes that will have a positive societal impact by design and by default.

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. Xenia Ziouvelou chaired the Ethicomp 2025 track "How is AI Changing Us? How Does it Change the World?" and was excluded from the review process of this paper.

References

1. Ghallab, M.: Responsible AI: requirements and challenges. AI Perspectives 1(1), 1–7 (2019).
2. Dove, E.: Research Ethics Review. In G. Laurie, E. Dove, A. Ganguli-Mitra, C. McMillan, E. Postan, N. Sethi, & A. Sorbie (Eds.), The Cambridge Handbook of Health Research Regulation, 177–186. Chapter, Cambridge University Press (2021).
3. 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 Medical Ethics 9, 6 (2008).
4. Page, S. A., & Nyeboer, J.: Improving the process of research ethics review. Research Integrity and Peer Review 2, 14 (2017).
5. Ferretti, A., Ienca, M., Sheehan, M., Blasimme, A., Dove, E. S., Farsides, B., Friesen, P., Kahn, J., Karlen, W., Kleist, P., Liao, S. M., Nebeker, C., Samuel, G., Shabani, M., Rivas Velarde, M., Vayena, E.: Ethics review of big data research: What should stay and what should be reformed? BMC Medical Ethics 22, Article 51 (2021).
6. Kitchenham, B., & Charters, S. M.: Guidelines for Performing Systematic Literature Reviews in Software Engineering. EBSE (2007).
7. Patton, M. Q.: Qualitative Evaluation and Research Methods (2nd ed.). Newbury Park, CA: Sage Publications (1990).
8. Lanzerath, D.: Research Ethics and Research Ethics Committees in Europe. In Zima, T., & Weisstub, D. N. (eds.), Medical Research Ethics: Challenges in the 21st Century. Philosophy and Medicine, vol 132. Springer, Cham (2023).
9. Bouhouita-Guermech, S., Gogognon, P., & Bélisle-Pipon, J. C.: Specific challenges posed by artificial intelligence in research ethics. Frontiers in Artificial Intelligence 6, 1149082 (2023).
10. 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, 121–136. Chapter, Routledge, London (2014).
11. Friesen, P., Yusof, A. N. M., & Sheehan, M.: Should the Decisions of Institutional Review Boards Be Consistent? Ethics & Human Research 41(4), 2–14 (2019).
12. Flores-Viva, J. M., & García-Peñalvo, F. J.: Reflections on the Ethics, Potential, and Challenges of Artificial Intelligence in the Framework of Quality Education (SDG4). Comunicar: Media Education Research Journal 31(74), 35–44 (2023).
13. Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., ... & Deveci, M.: A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion 96, 156–191 (2023).
14. 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. Machine Learning and Knowledge Extraction 5(3), 1023–1035 (2023).
15. 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 Proceedings: Digital Health 1(2), 120–138 (2023).
16. Kumar, P., Chauhan, S., & Awasthi, L. K.: Artificial intelligence in healthcare: review, ethics, trust challenges & future research directions. Engineering Applications of Artificial Intelligence 120, 105894 (2023).
17. 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 eXtend-

ed Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), 723–728. IEEE (2023, October).
18. 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).

WHY HEMISPHERE THEORY MATTERS FOR HOW WE THINK ABOUT ETHICS AND AI IN MEDICINE

HUGH BROSNAHAN[1] [0009-0004-8305-4852]

Keywords: Artificial Intelligence, Clinical Ethics, Hemisphere Theory, McGilchrist, Mumford, AI scribes

Abstract

The use of artificial intelligence (AI) in medicine risks reinforcing a paradigm that prioritises efficiency and protocol adherence over expert intuition and patient-centred care. Drawing on Iain McGilchrist's theory of hemispheric specialisation, this paper argues that AI systems, such as scribes, represent an augmented left-hemispheric orientation—favouring abstraction, categorisation, and rule-based processing at the expense of the right hemisphere's holistic, context-sensitive, and relational modes of engagement. The adoption of AI scribes in healthcare is often framed as a solution to administrative burdens, yet their integration is part of a deeper shift in how medical expertise is conceptualised and exercised. These technologies carry ethical, epistemological, and cognitive implications that extend beyond concerns of privacy, security, and algorithmic bias. This paper explores how the dominance of left-hemispheric cognition, augmented by AI integration, jeopardises the moral and relational dimensions of clinical care—areas where the right hemisphere's contributions to empathy, moral reasoning, and context-sensitive judgment are vital. Framed within Lewis Mumford's concept of the "Machine"—the reductionistic organisation of society around principles of efficiency, standardisation, and control—AI is viewed as actively reinforcing a model of healthcare that undermines the human dimensions of medical practice.

Introduction

The integration of artificial intelligence (AI) into clinical settings is not just a technical innovation but emblematic of a broader societal shift in which technology increasingly mediates human interaction and decision-making (e.g., Coeckelbergh, 2020; Haidt, 2024; Rosen, 2024). AI scribes, designed to record, transcribe, and summarise patient- clinician interactions, are often touted as solutions to administrative overload, ostensibly freeing healthcare professionals to focus on patient care (Mess et al., 2025). However, their adoption raises ethical, epistemological, and practical concerns that ex

1 UNIVERSITY OF OTAGO, DUNEDIN 9016, NEW ZEALAND, HUGH.BROSNAHAN@POSTGRAD.OTAGO.AC.NZ

tend beyond issues of privacy, security, and bias (e.g., Cresswell et al., 2024; Scassa & Kim, 2025). This paper argues that an underexamined dimension of AI's role in healthcare is its alignment with the left hemisphere's mode of attention and the resulting impact on how we understand and practice medicine.

Theoretical Framework

Adopting Iain McGilchrist's (2009, 2021) theory of hemispheric specialisation, this paper argues that AI systems, including scribes, are a radical expression of left-hemispheric attention, privileging abstraction, categorisation, and rule-based processing over the right hemisphere's holistic, context-sensitive, embodied, and intuitive engagement with the world. McGilchrist's work distinguishes the hemispheres less by what they do than how they do it, revealing two ultimately incommensurate worldviews. The left hemisphere allows us to apprehend aspects of reality schematically, isolating parts from the whole for the purposes of conceptualisation, manipulation and control, whereas the right hemisphere enables relational comprehension of the world as a meaningful, dynamic, and inexhaustible whole. Crucially, the left hemisphere's fragmented vision is dependent on the right, producing useful but necessarily reductive "maps" of the "terrain" the right hemisphere more directly encounters. In this framework, the boundaries of language—understood as any system of signs used to represent and manipulate information—are coextensive with the left hemisphere's cognitive paradigm, suggesting the inherent limitations of AI (Kastrup, 2016; Jaeger et al., 2024). AI systems are fundamentally constrained by their inability to grasp meaning the way that human beings do. They process language algorithmically using statistical associations between symbols (resembling a left-hemispheric approach to language), rather than experientially to intuit meaning, generating responses that appear coherent but lack the understanding that comes from embodied understanding (Jaeger et al., 2024; McGilchrist, 2021, Vigneau et al., 2011). Unlike human minds, where the right hemisphere counterbalances the left's reductive tendencies, AI lacks this corrective function, depending on human users to provide contextualisation and relevance realisation (Vervaeke et al., 2009). Viewing these technologies through the lens of Lewis Mumford's (1967) concept of the "Machine"—the vast, systemic organisation of society, structured around principles of efficiency, standardisation, and control, enabled and reinforced by technology—AI scribes are not neutral tools but active reinforcers of this paradigm. On this view, their integration into clinical practice serves to further entrench a left-hemispheric vision of medicine, where protocol adherence, data optimisation, and measurable outcomes take precedence over the tacit knowledge, clinical intuition, and relational understanding that constitute the art of medical practice.

The Cognitive and Ethical Implications of AI in Medicine

The nature of hemispheric lateralisation is not merely a matter of cognitive preference but has profound implications for how clinicians engage with patients, interpret medical data, and make ethical decisions. McGilchrist (2021) highlights how

the right hemisphere plays a critical role in assessing the 'morality of actions, moral reasoning, and promoting prosocial norms' (p. 1343), a view supported by multiple studies (e.g., Hecht, 2014; Kuhl & Kazén, 2008; Miller et al., 2010; Mychack et al., 2001; Schultheiss, 2018). Delusion, confabulation, unwarranted optimism, and an unwillingness to reconsider hypotheses in the face of conflicting evidence are further aspects of left-hemispheric attention (e.g., Marinsek et al., 2014; McGilchrist, 2021; Sharot et al., 2012). If AI scribes and decision-support systems become a dominant presence in the clinical environment, we risk not only reinforcing left-hemispheric dispositions—prioritising efficiency, standardisation, and protocol adherence at the expense of intuition, empathy, and relational understanding—but also succumbing to rigid, flawed and destructively idealistic frameworks that resist correction, elevating models over the reality they seek to represent.

A key promise of AI scribes is their potential to restore the doctor-patient relationship by reducing the cognitive and administrative overload imposed by electronic health records (EHRs) (e.g., Ma et al., 2025; Tierney et al., 2024). However, we must weigh this benefit against the possibility of clinicians becoming overly reliant on these systems, resulting not only in automation bias (Mehta & Mehta, 2024; Nguyen, 2024) but potentially leading to a gradual erosion of their own ability to engage in the kind of patient-centred and experience-derived reasoning that characterises true medical expertise (Dreyfus & Dreyfus, 1986).

The introduction of AI into clinical settings risks exacerbating a broader epistemic shift that has been underway for decades—a shift toward construing knowledge as primarily explicit, quantifiable, and rule-governed, rather than as something that has implicit, experiential, and contextual understanding as its foundation (Dreyfus & Dreyfus, 1986; McGilchrist, 2009, 2021; Mumford, 1967). This shift is particularly concerning in healthcare, where effective practice depends not only on the application of protocols and evidence-based guidelines but also on the clinician's ability to read subtle cues, synthesise disparate sources of information, and navigate the complex, often ambiguous reality of human health (Ambady et al., 2002; Bateson, 1973; Hong & Oh, 2020; Lee et al., 2019; Levinson et al., 1997).

If AI systems reinforce a mode of clinical reasoning that is predominantly left- hemispheric in nature, detached from the experience of both patients and practitioners (Shin et al., 2024), we risk diminishing the moral and relational dimensions of medicine. Ethical judgement in healthcare has traditionally drawn from deontological, consequentialist, and virtue-ethics frameworks. However, AI, insofar as it mirrors left-hemispheric conceptions of morality, aligns most naturally with explicit, procedural, and quantifiable modes of thought—characteristics shared by both rule-based (deontological) and cost-benefit (utilitarian) theories (Dreyfus & Dreyfus, 1986; McGilchrist, 2009, 2021). While these approaches have their place in medical ethics, sidelining the intuitive and relational dimensions of moral reasoning, which are central to virtue ethics, will result in the predominance of calculative rationality at the expense of true ethical practice.

Conclusion

To mitigate these challenges, AI integration in healthcare must be approached with a clear recognition of its cognitive and ethical implications. AI-driven clinical practice may serve to entrench a left-hemispheric mode of reasoning that prioritises explicit rules, efficiency, and abstraction over the intuitive, contextual, and meaning-laden dimensions of clinical expertise (Dreyfus & Dreyfus, 1986; McGilchrist, 2009, 2021). The future of AI in healthcare should not be reduced—left hemisphere style—to a binary choice between wholesale adoption or outright rejection. Rather than allowing AI to redefine the practice of medicine in its own image, its role should be that of a tool that augments rather than supplants the experiential foundations of good clinical judgment. Cultivating the art of medicine requires that clinicians remain attuned to the implicit and embodied dimensions of care that AI cannot replace, ensuring that ethical and clinical wisdom are not reduced to algorithmic processes. By integrating hemisphere theory into AI ethics, we can steer technological development toward an approach that respects the primacy of human understanding, the richness of experience, and the necessary depth demanded of patient care.

Acknowledgments: The author's research is funded by the University of Otago Doctoral Scholarship.

Disclosure of Interests: The author declares no conflict of interest.

References

1. Ambady, N., LaPlante, D., Nguyen, T., Rosenthal, R., Chaumeton, N., & Levinson, W. (2002). Surgeons' tone of voice: A clue to malpractice history. Surgery, 132(1), 5–9. https://doi.org/10.1067/msy.2002.124733
2. Coeckelbergh, M. (2020). AI Ethics. MIT Press.
3. Cresswell, K., Keizer, N. de, Magrabi, F., Williams, R., Rigby, M., Prgomet, M., Kukhareva, P., Wong, Z. S.-Y., Scott, P., Craven, C. K., Georgiou, A., Medlock, S., McNair, J. B., & Ammenwerth, E. (2024). Evaluating Artificial Intelligence in Clinical Settings—Let Us Not Reinvent the Wheel. Journal of Medical Internet Research, 26(1), e46407. https://doi.org/10.2196/46407
4. Dreyfus, H. L., & Dreyfus, S. E. (1986). Mind Over Machine: The Power of Human Intuition and the Era of the Computer. Basil Blackwell.
5. Haidt, J. (2024). The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness. Penguin Press.
6. Hecht, D. (2014). Cerebral Lateralization of Pro- and Anti-Social Tendencies. Experimental Neurobiology, 23(1), 1–27. https://doi.org/10.5607/en.2014.23.1.1
7. Hong, H., & Oh, H. J. (2020). The Effects of Patient-Centered Communication: Exploring the Mediating Role of Trust in Healthcare Providers. Health Communication, 35(4), 502–511. https://doi.org/10.1080/10410236.2019.1570427
8. Jaeger, J., Riedl, A., Djedovic, A., Vervaeke, J., & Walsh, D. (2024). Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. Frontiers in Psychology, 15. https://doi.org/10.3389/fpsyg.2024.1362658
9. Kastrup, B. (2016). More Than Allegory: On Religious Myth, Truth and Belief. iff Books.
10. Kuhl, J., & Kazén, M. (2008). Motivation, affect, and hemispheric asymmetry: Power versus affiliation. Journal of Personality and Social Psychology, 95(2), 456–469. https://doi.org/10.1037/0022-3514.95.2.456

11. Lee, M., Murphy, K., & Andrews, G. (2018). Using Media While Interacting Face-to-Face Is Associated With Psychosocial Well-Being and Personality Traits. Psychological Reports, 122(3), 944–967. https://doi.org/10.1177/0033294118770357
12. Levinson, W., Roter, D. L., Mullooly, J. P., Dull, V. T., & Frankel, R. M. (1997). Physician-Patient Communication: The Relationship With Malpractice Claims Among Primary Care Physicians and Surgeons. JAMA, 277(7), 553–559. https://doi.org/10.1001/jama.1997.03540310051034
13. Ma, S. P., Liang, A. S., Shah, S. J., Smith, M., Jeong, Y., Devon-Sand, A., Crowell, T., Delahaie, C., Hsia, C., Lin, S., Shanafelt, T., Pfeffer, M. A., Sharp, C., & Garcia, P. (2025). Ambient artificial intelligence scribes: Utilization and impact on documentation time. Journal of the American Medical Informatics Association, 32(2), 381–385. https://doi.org/10.1093/jamia/ocae304
14. Marinsek, N., Turner, B. O., Gazzaniga, M., & Miller, M. B. (2014). Divergent hemispheric reasoning strategies: Reducing uncertainty versus resolving inconsistency. Frontiers in Human Neuroscience, 8, 839. https://doi.org/10.3389/fnhum.2014.00839
15. McGilchrist, I. (2010). The Master and His Emissary: The Divided Brain and the Making of the Western World. Yale University Press.
16. McGilchrist, I. (2021). The Matter With Things: Our Brains, Our Delusions, and the Unmaking of the World: Vol. I: The Ways to Truth. Perspectiva Press.
17. Mehta, S., & Mehta, N. (2024). Embracing the illusion of explanatory depth: A strategic framework for using iterative prompting for integrating large language models in healthcare education. Medical Teacher, 1–4. https://doi.org/10.1080/0142159X.2024.2382863
18. Mess, S. A., Mackey, A. J., & Yarowsky, D. E. (2025). Artificial Intelligence Scribe and Large Language Model Technology in Healthcare Documentation: Advantages, Limitations, and Recommendations. Plastic and Reconstructive Surgery – Global Open, 13(1), e6450. https://doi.org/10.1097/GOX.0000000000006450
19. Miller, M. B., Sinnott-Armstrong, W., Young, L., King, D., Paggi, A., Fabri, M., Polonara, G., & Gazzaniga, M. S. (2010). Abnormal moral reasoning in complete and partial callosotomy patients. Neuropsychologia, 48(7), 2215–2220. https://doi.org/10.1016/j.neuropsychologia.2010.02.021
20. Mychack, P., Kramer, J. H., Boone, K. B., & Miller, B. L. (2001). The influence of right frontotemporal dysfunction on social behavior in frontotemporal dementia. Neurology, 56(suppl_4), S11–S15. https://doi.org/10.1212/WNL.56.suppl_4.S11
21. Nguyen, T. (2024). ChatGPT in Medical Education: A Precursor for Automation Bias? JMIR Medical Education, 10(1), e50174. https://doi.org/10.2196/50174
22. Rosen, C. (2024). The Extinction of Experience: Being Human in a Disembodied World. W. W. Norton & Company, Inc.
23. Scassa, T., & Kim, D. (2025). AI Medical Scribes: Addressing Privacy and AI Risks with an Emergent Solution to Primary Care Challenges
 (SSRN Scholarly Paper No. 5086289). Social Science Research Network. https://papers.ssrn.com/abstract=5086289
24. Schultheiss, O. C. (2018). Implicit motives and hemispheric processing differences are critical for understanding personality disorders: A Commentary on Hopwood. European Journal of Personality, 32(5), 580-582.
25. Sharot, T., Kanai, R., Marston, D., Korn, C. W., Rees, G., & Dolan, R. J. (2012). Selectively altering belief formation in the human brain. Proceedings of the National Academy of Sciences, 109(42), 17058–17062. https://doi.org/10.1073/pnas.1205828109
26. Shin, M., Taseski, D., & Murphy, K. (2024). Media multitasking is linked to attentional errors, mind wandering and automatised response to stimuli without full conscious processing. Behaviour & Information Technology, 43(3), 445–457. https://doi.org/10.1080/0144929X.2023.2167669
27. Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Kipnis, P., Liu, V., & Lee, K. (2024). Ambient Artificial Intelligence Scribes to Alleviate the Burden of

28. Clinical Documentation. NEJM Catalyst, 5(3), CAT.23.0404. https://doi.org/10.1056/CAT.23.0404
29. Vervaeke, J., Lillicrap, T. P., & Richards, B. A. (2009). Relevance Realization and the Emerging Framework in Cognitive Science. Journal of Logic and Computation, 22(1), 79–99. https://doi.org/10.1093/logcom/exp067
30. Vigneau, M., Beaucousin, V., Hervé, P.-Y., Jobard, G., Petit, L., Crivello, F., Mellet, E., Zago, L., Mazoyer, B., & Tzourio-Mazoyer, N. (2011). What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing?: Insights from a meta-analysis. NeuroImage, 54(1), 577–593. https://doi.org/10.1016/j.neuroimage.2010.07.036

HOW AI IS RESHAPING CREATIVITY: DEEPSEEK VS CHATGPT PLUS IN LEGO® SERIOUS PLAY®

ANTÓNIO ALMEIDA[1], DENISE MEYERSON[2], INÊS ALMEIDA[3]

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

Extended Abstract

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

1 INSTITUTO SUPERIOR DE CIÊNCIAS SOCIAIS E POLÍTICAS DA UNIVERSIDADE DE LISBOA, PORTUGAL
2 UNIVERSITY OF THE WITWATERSRAND, SOUTH AFRICA
3 AUTÓNOMA TECHLAB, UNIVERSIDADE AUTÓNOMA DE LISBOA, PORTUGAL; INSTITUTO DE TELECOMUNICAÇÕES, PORTUGAL; FCT, NOVA UNIVERSITY OF LISBON, CAPARICA PORTUGAL

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 [13]. 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 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 study also incorporates insights from AI research methodologies and structured prompt design approaches, such as those proposed by Ethan Mollick at the Wharton School, who has explored structured prompt engineering to optimize generative AI outputs [12]. By incorporating best practices in prompt structuring, this study seeks to refine the AIs ability to produce session plans that are more aligned with LEGO® SERIOUS PLAY® facilitation practices, ensuring that comparisons between AI generated and human-generated plans are as precise as possible (see Fig.1).

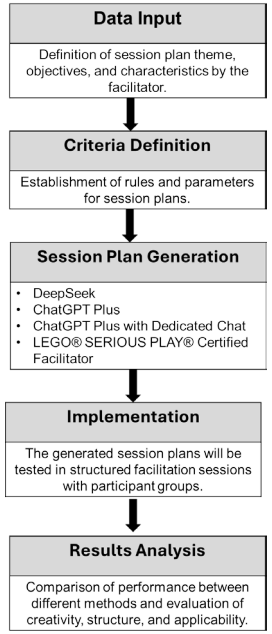


Fig1. Research Methodology for AI-Generated LEGO® SERIOUS PLAY® Session Plans.

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].

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 Education 10(1), 2262284 (2023)
3. Roos, J., Victor, B.: How it all began: The origins of LEGO® Serious Play®. International Journal of Management and Applied Research 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®. European Journal of Innovation Management 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. Marketing Education Review 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. The Wharton School Research Paper (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. arXiv preprint arXiv:2501.12948 (2025)

ARTIFICIAL INTELLIGENCE INTEGRATION IN EVERY LIFE – UTOPIAN FRIEND OR DYSTOPIAN FOE?

Minna M. Rantanen[1] [0000-0001-8832-5616], Anushree Luukela-Tandon[2] [0000-0002-7319-418X], Salla Westerstrand[3] [0000-0002-1441-8673], Jani Koskinen[4] [0000-0001-8325-9277]

Keywords: Artificial intelligence, Lifeworld, Systemic powers, Discourse ethics, Focus group

Extended Abstract

Information Systems (ISs) like Artificial Intelligence (AI) have become increasingly pervasive in human lives. We use AI tools such as Alexa or Siri for daily tasks, view content recommended by an algorithm, and use healthcare apps to assess symptoms before obtaining expert human care. We use AI tools to expand bullet points and summarise large masses of text, curate our ideas and make sense of those of others. Such systems have become an integral part of many human processes as they "penetrate human life, experiences, products, business processes and civil society" [1 p. 47]. This degree of AI-human intertwining has been noted by scholars and practitioners alike – both communities are focused on debating the pros and cons of this phenomenon, including the definition of what progress and good life should look like for everyone.

Practitioners often applaud the efficiency of using AI for reasons such as improved business operations and garnering consumers' attention. For example, recently, IAB [2] surveyed advertising professionals in Europe about their thoughts on the value AI brought to their business and found an overwhelming majority to concede for positive effects like gaining operational efficiencies (78%) and competitive advantage (60%). Similarly Silo AI [3, 4] reported that over 77% of Nordic organisations actively use Generative AI, while only about 2% reported using no AI technology. The same poll also found that over 70% of participating organisations prioritised AI development for the coming year.

Businesses' prioritisation of AI in their operations seems supported by consumers' primarily positive response to its benefits. For instance, Frank et al. [5] determined that consumer demand for AI-assisted products stemmed from their acknowledgement of its utilitarian, hedonic, and symbolic value, contingent factors such as national culture and individual self-construal. In another study, Chakraborty et al. [6] determined that consumers' willingness to adopt AI-assisted experience in the beauty and cosmetics industry

1 University of Jyväskylä, Finland, minna.m.rantanen@jyu.fi

2 University of Eastern Finland, Finland; European forest institute, Finland

3 University of Turku, Finland

4 University of Turku, Finland

was moderated by the virtual trial experience possibility and significantly affected by a cascading effect of variables such as AI compatibility, perceived performance expectancy and perceived ease of use.

Meanwhile, a body of research has highlighted the negative side of the pervasive use of AI and emphasised its ability to promulgate a surveillance culture [7–9], bias and discrimination [10–12] and impede freedom and human autonomy [13]. For example, Stahl and Eke [14] studied the ethical impacts of ChatGPT to assess the benefits and risks that the tool can surface and identified eleven benefits (e.g., human identity, labour market, sustainability and health) and thirty ethical concerns (e.g., psychological harm, harms to society, environmental harm, and autonomy). Such studies indicate that the ethics of using AI to add convenience to business and consumer activities is a broad topic with overlapping positives and negatives that scholars are debating.

However, even scholars may be privy to the ethical issues stemming from using AI. In a recent scoping review, Bouhouita-Guermech et al. [15] highlighted that the evolution of AI for supporting research activities have spawned specific ethical challenges, including the lack of AI-specific standards and governance structures to support research ethics.

Thefore, it seems evident that the pervasiveness of AI into human lives is a topic that needs more research to be properly understood. However, how is the pervasiveness of AI perceived by individuals? Does an individual stop to think how and when they should use AI, how AI providers use their data, and what that means to their privacy and autonomy– not to mention the ethical justifications of their perceptions? According to consumer statistics, the answer is ambiguous. To exemplify, a poll among United States consumers [16] reported that most respondents (88% Gen Z and 77% millennials) expected AI to improve their online shopping experiences. In another survey, about 47% of Japanese consumers expressed that Gen AI could be optimally used to enhance consumer services [17]. However, another poll among Spaniards [18] found 74% of respondents to agree that AI was evolving at a rate faster than society's ability to assimilate it, and 44% distrusted people's ethical use of this technology. Thus, there is no clear answer to how such pervasive AI use is perceived or about the ethics of the pervasiveness and its expected impacts.

This study aims to fill this gap and increase understanding of individuals' perceptions of pervasiveness of AI systems. We take a critical approach and use a focus group as a method to map how people see the role of AI systems in different dimensions of their lifeworld and their attitudes towards those roles. Focus group has been shown to be suitable for critical IS research [19], which makes it a viable approach for profoundly understanding people's perception of pervasive AI systems and their ethical implications. We follow the definition of Powell and Single [20 p. 499], according to whom a focus group is "a group of individuals selected and assembled by researchers to discuss and comment on, from personal experience, the topic that is the subject of the research." It is particularly suitable for complex research topics that benefit from additional validation and sensemaking with stakeholders [20] , which is also the case of the present study. To gain both a general image of perceptions and in-depth knowledge on their grounds, this study includes five focus groups focusing on different areas of life where AI has gained increasing attention: academic use of AI (knowledge creation), private sector

use of AI (organisational context), and individual use of AI in everyday life. Each focus group consists of 5-7 people from different backgrounds to achieve varied perspectives of stakeholders.

To ensure the analysis is ethically solid, we use Habermas's lifeworld/system -model [21–23] to conceptualise the ethics of the pervasiveness of AI systems in human life. In this model, lifeworld is a sphere where individuals share a common reality, encounter each other and communicate in everyday life. In the lifeworld, communication (ideally) leads to mutual understanding and consensus-seeking discourse with others, i.e., communicative action. In contrast, systems refer to economic, political, administrative and technological systems driven by rationalities that serve the needs and aims of those institutional systems rather than specific individuals.

Problems arise when systemic powers colonise the lifeworld and implement rationalities of the systems into lifeworlds. For example, the institutional logics and value systems embedded in AI algorithms are not those of their users but of the systemic powers, defining, e.g., the rules and the infrastructure of digital platforms we communicate on.

While systemic powers can also cultivate the essence of our lifeworlds, taking systemic rationalities as given risks making them the driving force of our lifeworld and thus colonising it. For instance, mobile phones and the systems behind them make it possible to connect people even if they are living in different locations. However, when the use of mobile phones is driven by the interest of the business that tries to make us spend as much time staring at our phones, depriving us of connection with others, we witness colonisation of our lifeworld by technology. Likewise, AI is loaded with big expectations and promises. However, the outcome can be a situation where AI colonises our lifeworlds and hence, the rationalities behind AI will be integrated into our lives [see 1].

Habermas's model offers an illustrative approach to understanding the pervasiveness of AI systems, as it describes rationalities, communication and coordination between lifeworlds and systemic powers. It offers a moral-philosophically sound foundation for a focus group study because it views discourse as an essential building block of ethical action. According to Habermas, ethical justifications are achieved through proceduralist, deliberative and communicative action that aims towards mutual understanding rather than mere strategic goals [23–25].

The strength of this approach is discourse ethics, which does not take a rigid standpoint characteristic to the normative ethical theories (deontology, consequentialism and virtue ethics). Instead, it accommodates different ethical viewpoints in discourse, serving as an arena for ethical justifications and perceptions. Therefore, it provides a mechanism that allows consideration of different moral views and intuitions, as it offers various tools for in-depth analyses of communication and discourse, making both past and current communication problems visible [26]. This approach has been used by, e.g., Rehg [27], who developed a framework for moral inquiry and reflection on problematic cyberpractices, and Yetim [28], who provided metacommunication model -- used for analysing communication (see also [29], [30]). Overall, Habermas' critical social theory, which we use here, has been one of the factors contributing to paradigm change in IS research [see 31–33]. Focus groups used in this study bring together people from different perspectives, which ensures comprehensive stakeholder inclusion – a prerequisite for communicative action and discourse ethics [34].

This study is thus a contribution to increasing our understanding of the ethics of pervasive digital systems and how people perceive the pervasiveness in real life. Hence, this combination is rich in theoretical and empirical insights that can be used to inform the design and governance of digital systems so that they bring ethical value to organisations and individuals.

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

References

1. Grover, V., Lyytinen, K.: The pursuit of innovative theory in the digital age. Journal of Information Technology. 38, 45–59 (2023).
2. Wakefield, L.: IAB Europe and Microsoft Advertising's Understanding the Adoption and Application of AI in Digital Advertising Report - IAB Europe, https://iabeurope.eu/knowledge_hub/iab-europe-and-microsofts-understanding-the-adoption-and-application-of-ai-in-digital-advertising-report/, last accessed 2025/02/06.
3. Silo AI: Nordic State of AI, third edition, https://www.silo.ai//ebooks-reports/nordic-state-of-ai-third-edition, last accessed 2025/02/06.
4. Silo AI: Artificial intelligence and the technologies used in 2023 by Nordic organizations. (November 1, 2023) In Statista. Retrieved February 06, 2025, from https://www.statista.com/statistics/1454469/ai-technologies-used-by-nordics/
5. Frank, B., Herbas-Torrico, B., Schvaneveldt, S.J.: The AI-extended consumer: Technology, consumer, country differences in the formation of demand for AI-empowered consumer products. Technological Forecasting and Social Change. 172, (2021). https://doi.org/10.1016/j.techfore.2021.121018.
6. Chakraborty, D., Polisetty, A., G, S., Rana, N.P., Khorana, S.: Unlocking the potential of AI: Enhancing consumer engagement in the beauty and cosmetic product purchases. Journal of Retailing and Consumer Services. 79, 103842 (2024). https://doi.org/10.1016/j.jretconser.2024.103842.
7. Zuboff, S.: Big other: Surveillance Capitalism and the Prospects of an Information Civilization. Journal of Information Technology. 30, 75–89 (2015). https://doi.org/10.1057/jit.2015.5.
8. Zuboff, S.: The Age of Surveillance Capitalism. Profile Books, London, England (2019).
9. Muldoon, J., Graham, M., Cant, C.: Feeding the machine: the hidden human labour powering AI. Canongate Books (2024).
10. Capitol Technology University: The Ethical Considerations of Artificial Intelligence, https://www.captechu.edu/blog/ethical-considerations-of-artificial-intelligence, last accessed 2025/02/06.
11. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR). 54, 1–35 (2021).
12. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K.: Algorithmic fairness: Choices, assumptions, and definitions. Annual review of statistics and its application. 8, 141–163 (2021).
13. Prunkl, C.: Human autonomy at risk? An analysis of the challenges from AI. Minds and Machines. 34, 26 (2024).
14. Stahl, B.C., Eke, D.: The ethics of ChatGPT–Exploring the ethical issues of an emerging technolo-gy. International Journal of Information Management. 74, 102700 (2024).
15. Bouhouita-Guermech, S., Gogognon, P., Bélisle-Pipon, J.-C.: Specific challenges posed by artificial intelligence in research ethics. Frontiers in Artificial Intelligence. 6, (2023). https://doi.org/10.3389/frai.2023.1149082.
16. Rokt, The Harris Poll: AI and the rise of relevancy: Shifting expectations of the modern shopper, https://get.rokt.com/harris-poll.

17. Ministry of Internal Affairs and Communications: Attitudes on the spread of generative artificial intelligence (AI) in Japan as of February 2024 [Graph]. In Statista. Retrieved February 06, 2025, from https://www-statista-com.ezproxy.uef.fi:2443/statistics/1546134/japan-attitudes-on-spread-of-generative-artificial-intelligence/.
18. AIMC: Points of view on artificial intelligence (AI) in Spain in 2023 [Graph]. (2024). [Graph]. In Statista. Retrieved February 06, 2025, from In Statista. Retrieved February 06, 2025, from https://www-statista-com/statistics/1490271/spain-opinions-about-artificial-intelligence/
19. Stahl, B.C., Tremblay, M.C., LeRouge, C.M.: Focus groups and critical social IS research: how the choice of method can promote emancipation of respondents and researchers. European Journal of Information Systems. 20, 378–394 (2011).
20. Powell, R.A., Single, H.M.: Focus groups. International journal for quality in health care. 8, 499–504 (1996).
21. Habermas, J.: The theory of communicative action, Volume 1. Polity Press (1984).
22. Habermas, J.: The Theory of Communicative Action: Lifeworld and Systems, a Critique of Functionalist Reason, Volume 2. Polity Press (1987).
23. Habermas, J.: Between facts and norms: Contributions to a discourse theory of law and democracy. Mit Press (1996).
24. Habermas, J.: Moral Consciousness and Communicative Action. Polity Press (1990).
25. Habermas, J.: Justification and Application. Polity Press (1993).
26. Stahl, B.C.: Morality, ethics, and reflection: a categorization of normative IS research. Journal of the association for information systems. 13, 1 (2012).
27. Rehg, W.: Discourse ethics for computer ethics: a heuristic for engaged dialogical re-flection. Ethics and Information Technology. 17, 27–39 (2015).
28. Yetim, F.: Acting with genres: discursive-ethical concepts for reflecting on and legitimat-ing genres. European Journal of Information Systems. 15, 54–69 (2006).
29. Ulrich, W.: Critical Systemic Discourse. Journal of Information Technology Theory and Application (JITTA). 3, 10 (2001).
30. Stansbury, J.: Reasoned moral agreement: Applying discourse ethics within organizations. Business Ethics Quarterly. 19, 33–56 (2009).
31. Lyytinen, K., Hirschheim, R.: Information systems as rational discourse: An application of Habermas's theory of communicative action. Scandinavian Journal of Management. 4, 19–30 (1988).
32. Hirschheim, R., Klein, H.K.: Four paradigms of information systems development. Communications of the ACM. 32, 1199–1216 (1989). https://doi.org/10.1145/67933.67937.
33. Ross, A., Chiasson, M.: Habermas and information systems research: New directions. Information and Organization. 21, 123–141 (2011). https://doi.org/10.1016/j.infoandorg.2011.06.001.
34. Tuikka, A.-M., Koskinen, J., Kimppa, K.K.: Habermasian discourse on digitalisation of governmental services–testing discourse ethical framework for communication. In: ICIS 2022 Proceedings (2022).

HOW AI MAY AFFECT OUR THINKING

Iordanis Kavathatzopoulos[1] [0000-0003-3806-5216]

Keywords: Artificial Intelligence, Thinking, Rationality, Heuristics.

Extended Abstract

Focus on AI and its users

Currently the risks of Artificial Intelligence (AI) are discussed intensively. This discussion focuses mainly on technical aspects or on internal technical processes which may give AI unforeseen super powers. The fears are that AI may become more intelligent than humans, emerge as an independent agent with its own agenda, force us into something we do not want to be, lead evolution in a radically different direction, and possibly transform the whole universe (see for example: [2] [3] [4] [8] [11]). Although we cannot to rule out these risks, and because we want to do something to eliminate the risks, we would need to look to other conditions than the exclusively technical features, which may be playing an important role in causing these concerns about a very powerful AI.

Let us first see what we expect from AI, and what AI actually does. AI, as it is designed and used today, provides answers to problems and delivers products and services. All our efforts to develop AI tools and systems aim to support their ability to solve problems we humans have, and to satisfy our needs. All interest around AI is concentrated on achievement of goals, solutions of problems, and satisfaction of needs. This is what we expect AI to deliver and we design it accordingly.

If we now turn our focus on ourselves and try to see how we manage to handle our problems we can identify a certain tool we have for this purpose. It seems that thinking is the biological tool for achieving goals, solving problems, and satisfying needs [9]. However, this is not something we like so much to use. For us it is not easy to reason in a rational way, and it is even less easy to control the process of thinking even though we are conscious about the high value of right thinking. For us as persons it is necessary to have the skill to run a rational thinking process and the emotional strength to take responsibility for our own decisions. But if we try to reason in a rational way, we need to put great effort in this, concentrate our focus and spend a lot of energy. We have also to postpone making the decision and proceeding to action until we can come to a conclusion and find a suitable answer to the problem at hand. On top of these difficulties is the feeling of responsibility and the anxiety that follows with it when one knows that if something goes wrong there is no one else to blame but oneself.

1 Uppsala University, Lägerhyddsvägen 1, 752 37 Uppsala, Sweden, Iordanis.Kavathatzopoulos@it.uu.se

Rational thinking

Suppose now that we, despite all these difficulties and costs in energy and time, still make the effort to think rationally in order to find the answer to the problem we have in front of us. Will we achieve a satisfying result? The answer is unfortunately "maybe". So, the optimal solution cannot be guaranteed through the adoption of a rational, philosophical or scientific approach. Not in the real world, compared to an ideal or metaphysical world (see for example the discussion on the application of the categorical imperative in real life [6]). The fact is that in real world non-rational ways of handling are successful very often. This has been studied empirically in psychology of human cognition and it is the so-called heuristic [5] [10]. In philosophy we have the Aristotelian habits, which also function very well [1].

Why should we then adopt the rational way of handling our problems given all the difficulties and the fact that we may still come to a wrong conclusion and action? When we also know that not so autonomous ways of thinking have a great chance to give us a satisfying solution? So, most of the time we do not adopt rational thinking.

AI can strengthen this. We have already seen the effects of technology on our thinking. For example, during the early times of Internet in the world many of us were enthusiastic about its potential to give us access to all kinds of information, especially information contradictory to our beliefs and certainties. We were seriously expecting it to have a healthy impact on thinking. Unfortunately, that proved to be wrong. Instead of promoting doubt and dialog it allowed people to look for and connect with other like-minded people in order, not to examine and falsify their beliefs, but to confirm what they already are convinced is true. Another aspect worth consideration is that algorithms, e.g. in social media, suggest links which are supposed to be in accordance with what the user likes.

Design and application of AI

The design and use of AI follows the same path. We create it in order to give us answers, not to ask questions that dispute our wishes or beliefs. We expect it to give us the products or the services we need. We want AI as a tool for the satisfaction of our needs, as our thinking is, but a biological one. The key point here is that advanced AI is or will be an immensely better tool. Will it then replace thinking?

A great risk is that this will happen. Unless we change focus in the design and use of AI: from the answers, or products or services provided, to the process leading to these results. An AI designed to help us use our thinking in order to find the answers and to produce the things we need, is something different than using AI to deliver the answers and the products directly to us. We can imagine AI as a partner in an ongoing dialog [7].

First and foremost, we have to be highly conscious about the difficulty of such a plan. For example, we have already seen how Internet is used, meaning that the same may happen with the use of AI. But if we want to use AI as a support tool for our thinking, to help us run the process of thinking in the right way, we have first to ask the question

whether this is possible at all. Are we able to accept or at least tolerate something that disturbs our state of mind and makes us doubt our convictions? Especially when we know very well that this same tool can deliver to us the best possible answers to our problems, and give us satisfaction, harmony and happiness?

This seems to be a very difficult plan to design and apply. However, it is possible to think that a compromise may give it a bigger chance. A compromise that is in accordance with the real-life use of thinking. We are never perfect philosophers, scientists or democrats. We cut corners and we use heuristics, but still, we manage somehow to reach our goals and to satisfy our needs. Then, what about an AI that sometimes and when there is no other better alternative because of the high demands of the problem at hand, support us to think in a rational way? But also, an AI that most often support us in finding and using the best possible heuristics and habits?

Disclosure of Interests: The author has no competing interests to declare that are relevant to the content of this article.

References

1. Aristoteles: Ἠθικά Νικομάχεια [Nicomachean ethics]. Papyros (1975)
2. Bostrom, N.: Superintelligence: Paths, dangers, strategies. Oxford University Press (2014)
3. Future of Life Institute: Pause giant AI experiments: an open letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/, last accessed 2025/3/31 (2023)
4. Harrari, Y. N.: Homo Deus. Natur & Kultur (2015)
5. Kahneman, D., Slovic, P. and Tversky, A.: Judgment under uncertainty: Heuristics and biases. Cambridge University Press (1982)
6. Kant, I.: Grundläggning av sedernas metafysik [Foundations of the metaphysics of morals]. Daidalos (2006)
7. Kavathatzopoulos, I.: Artificial Intelligence and the sustainability of thinking: How AI may destroy us, or help us. In T. T. Lennerfors and K. Murata (Eds.), Ethics and Sustainability in Digital Cultures (pp. 19–30). Routledge (2024)
8. Kurzweil, R.: The Singularity is near: When humans transcend biology. Penguin Books (2006)
9. Platon: Πρωταγόρας [Protagoras]. Zacharopoulos (1986)
10. Sunstein, C.R.: Moral heuristics. Behavioral and Brain Sciences, 28(4), 531–542. https://doi.org/10.1017/S0140525X05000099 (2005)
11. Tegmark, M.: Life 3.0: Being human in the age of artificial intelligence. Knopf Publishing Group (2017)

CONSTRUCTING AI GOVERNANCE: A STUDY OF POLICY DEVELOPMENT IN THE HR SECTOR

FREDERIC HEYMANS[1]

Keywords: AI governance, HR, ethics, temporary employment agencies, HR services providers.

Abstract

This study explores the internal governance of Artificial Intelligence (AI) in the HR sector, focusing on temporary employment agencies and HR consulting firms. While AI technologies promise efficiency, automation, and data-driven decision-making, their implementation in HR also brings significant ethical challenges, particularly regarding fairness, transparency, and bias. Despite the growing adoption of AI, there is little research on how these organizations structure their internal governance around such technologies. This paper seeks to fill that gap by evaluating the presence and current state of internal AI policies within these firms. Specifically, it investigates how these businesses have developed and implemented policies to govern their use of AI technologies and assesses the perceived return on investment (ROI) of incorporating ethical considerations into AI practices. Additionally, the study identifies the key challenges and best practices in AI governance and looks at the role stakeholders play in shaping these policies. We approach the theme through the lens of neo-institutional theory. Subsequently by analyzing expert interviews and case study research, the findings are contextualized within broader regulatory discussions, including alignment with the European Union's AI Act. By bridging the gap between technological advancement and responsible implementation, this study contributes to the evolving discourse on AI ethics in HR

Introduction

In recent years, the rapid advancement and widespread adoption of AI have significantly impacted various economic sectors. The Human Resources (HR) sector is among those experiencing accelerated growth in the implementation of AI applications [1]. From automated resume screening and payroll processing to AI-driven candidate matching and chatbots for initial interviews, AI technologies promise efficiency and sound data-driven decision- making in HR processes [2], [3], [4]. AI can also provide valuable insight into employee engagement, performance and potential attrition, what can consequently help HR professionals to make data-driven decisions and implement targeted interventions. Other possibilities include skills gap analysis of employees and the use of speech and video recognition to evaluate video interviews.

1 IMEC-SMIT, VUB, 1050 BRUSSELS, BELGIUM.

More recently, the emergence of generative AI in the workplace has opened the door to numerous new applications. The integration of AI in HR is not merely a technological shift; it also represents a fundamental change in how organizations manage a very valuable and sensitive asset: human capital [5]. Beyond operational improvements, AI fundamentally alters the dynamics of decision-making in HR by introducing algorithms and data-driven processes that impact human lives directly. Recruitment, hiring, promotions, and even employee retention decisions—traditionally guided by human intuition, experience, and ethical considerations—are now increasingly influenced by AI systems. This shift not only changes the tools HR professionals use but also transform the way human capital is perceived and managed within organizations.

Despite the growing adoption of AI in HR, the risks associated with its use—particularly when mishandled—are substantial. These risks are especially pronounced in the HR sector because AI systems interact directly with human capital. For instance, AI systems trained on historical data may inadvertently perpetuate biases, as evidenced by Amazon's experimental AI recruiting tool, which exhibited bias against women [6]. The lack of transparency in AI decision-making, often referred to as the 'black box' problem, further complicates issues of accountability and trust [7]. Given these risks, it is important to examine how businesses, particularly HR consulting firms and temporary employment agencies, are managing the complexities of AI implementation.

Current research in AI focuses on principles and guidelines for AI in general while this study aims to investigate governance in practice; the extent to which policies are implemented [8], [9]. While existing studies have largely focused on large corporations that integrate AI into established HR processes, there is a gap in understanding how temporary employment agencies and HR consulting firms, which operate in more fluid and diverse environments, structure their AI-related policies [1]. These organizations interact with a broad range of industries and candidates, which makes them a fertile ground for AI applications in recruitment, screening, and matching. However, the distinct challenges they face—such as fast-paced decision-making, high candidate turnover, and legal constraints related to employment rights and fairness—amplify the risks associated with AI [10]. This paper aims to address this research gap by examining the internal policies of these firms regarding AI use. Understanding the extent to which these organizations have developed formal policies, ethical guidelines, and governance frameworks will provide insights into the broader landscape of AI in HR and help identify best practices and regulatory needs.

The content of these policies and frameworks—and how it is crafted—opens up a deeper area of exploration. It raises the question of whether companies approach policy development as a mere formality, a "checkbox" exercise to ensure legal compliance and satisfy external stakeholders, or whether they see it as a strategic investment with tangible returns. The depth and quality of such policies often depend on whether organizations view them as an opportunity for long-term value creation [11]. For this reason, in this article, we also examine how companies position their policies against broader regulatory and ethical discussions. For instance, experts indicate the implications of policy frameworks such as the EU AI Act. Besides alignment, we also look at how mature organizations are in evaluating themselves and trying to measure the impact of their policies.

This focus on maturity leads us to the practical challenges of policy implementation. Given the rapid evolution of AI and evolving regulatory frameworks, internal policies must remain dynamic and adaptable. Our research shows that organizations of all sizes encounter significant difficulties in this regard. Large, multinational organizations, struggle to maintain oversight while continuously recalibrating policies to accommodate the diverse needs (both politically and culturally) of their branches across different jurisdictions. Meanwhile, smaller organizations face classic constraints, such as limited human resources and expertise to respond to evolving policy requirements.

Additionally, the role of stakeholders in shaping these policies can be insightful. External stakeholders, such as regulators, customers, and industry partners, and internal stakeholders—ranging from executives to data scientists—may have differing perspectives on what a "good" AI policy looks like [12]. Thus, an interesting aspect to investigate is the extent to which stakeholders, both internal and external, influence policy content. Do businesses merely follow external pressure to appear compliant, or is there an active collaboration with stakeholders to create policies that reflect a genuine commitment to responsible AI governance? Understanding this dynamic is key to evaluating whether companies see these policies as merely procedural or as integral components of a broader ethical and operational strategy.

Method

This study adopts a qualitative exploratory approach. Qualitative methods allow for an in-depth examination of organizational policies, decision-making processes, and stakeholder perspectives. Expert interviews are used to capture rich, detailed insights into organizational practices and strategies. This is supplemented with two case studies to broaden the sample, with a primary focus on the implementation of AI applications and the evaluation of the internal governance surrounding them.

The primary target population comprises Belgian HR service providers and temporary employment offices that either (1) currently use AI tools (e.g., automated CV screening, chatbots for candidate interactions) or (2) are in the process of developing or implementing AI-driven systems. Participants should hold decision-making or advisory roles in relation to AI usage, such as Chief HR Officers, AI Ethics Officers, Compliance Officers, Chief Information Officers, or Senior Digital Strategists. A purposive sampling method was used, to ensure the selection of participants who are directly knowledgeable about AI governance. The target sample size will be approximately 15 interviews. For the case studies, two large HR service organizations were selected that already have a policy in place for the implementation and use of AI applications.

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

References

1. Bujold, A., Roberge-Maltais, I., Parent-Rocheleau, X., Boasen, J., Sénécal, S., & Léger, P. (2023). Responsible artificial intelligence in human resources management: a review of the empirical literature. AI And Ethics. https://doi.org/10.1007/s43681-023-00325-1
2. Bankins, S. (2021). The ethical use of artificial intelligence in human resource management: a decision-making framework. Ethics and Information Technology, 23(4), 841-854.

3. Budhwar, P., Malik, A., De Silva, M. T. T., & Thevisuthan, P. (2022). Artificial intelligence – challenges and opportunities for international HRM: a review and research agenda. International Journal Of Human Resource Management, 33(6), 1065–1097. https://doi.org/10.1080/09585192.2022.2035161
4. Kupfer, C., Prassl, R., Fleiß, J., Malin, C., Thalmann, S., & Kubicek, B. (2023). Check the box! How to deal with automation bias in AI-based personnel selection. Frontiers in Psychology, 14, 1118723.
5. Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. California Management Review, 61(4), 15-42.
6. Dastin, J. (2018). Insight—Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/
7. Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K., & Jiya, T. (2022). Organisational responses to the ethical issues of artificial intelligence. AI & SOCIETY, 37(1), 23-37.
8. Bar-Gil, O., Ron, T., & Czerniak, O. (2024). AI for the people? Embedding AI ethics in HR and people analytics projects. Technology in Society, 77, 102527. https://doi.org/10.1016/j.techsoc.2024.102527
9. Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda. Journal of Business Ethics, 178(4), 977–1007. https://doi.org/10.1007/s10551-022-05049-6
10. Schellmann, H. (2024). The algorithm. Hachette Books.
11. Bevilacqua, M., Berente, N., Domin, H., Goehring, B., & Rossi, F. (2023). The Return on Investment in AI Ethics: A Holistic Framework. arXiv preprint arXiv:2309.13057.
12. Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2022). Designing fair AI in human resource management: Understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions. In Proceedings of the 2022 CHI conference on human factors in computing systems (pp. 1-22).

TOO GOOD TO BE TRUE? BUSINESS STUDENTS' ETHICAL USE OF AI FOR RESEARCH – TODAY AND BEYOND

Dr. Martha J. Wilcoxson[1]

Keywords: AI Ethics Policy; Business Ethics; University Curriculum Development.

Extended Abstract

A struggling graduate student under my supervision submitted a stellar literature review. The language was not his usual style, the citations and references were abundant (and mostly accurate), and the conclusion was valid and insightful. The work was almost too good to be true. Like many other business school professors, I am increasingly faced with the challenge of recognizing submissions of text authored by AI and proving this authorship. I support students' use of AI tools in research projects. Business schools cannot afford to ignore the fact that AI marks a powerful evolution of our civilization [1], [2], [3]. However, as a professor of business ethics, I have become increasingly confounded about how to approach students' use of AI-generated submissions in a manner that supports critical learning as well as their personal business-ethical development.

For the educator, student use of AI to write submissions presents four primary challenges. First, student use of AI is nearly undetectable [4]. The conundrum is that the instructor may suspect—but cannot prove—an AI submission. The professor may have a sense of the individual student's writing style and ability, but they cannot conclusively identify AI-generated submissions or accuse the student of using AI. Given that the ability to detect AI will only become weaker as AI evolves for abstract reasoning, this weakness will likely become more problematic [5]. Second, a significant number of my students, like graduate students at all other business schools, use AI to write or help write academic papers [6]. It is not merely the few tech-savvy students who are using AI platforms to generate submissions. Institutions are not able to validate the percentage of use frequency, but they know that students have widespread access to platforms such as ChatGPT [6]. The third challenge is that the student may use and even require AI to build content but learn nothing in the process. Student needs include learning and improving skills, saving time, and receiving the highest grades possible [4]. The concern with students using AI to complete academic work is that students can submit work and receive high marks without learning anything. Finally, AI platforms have the potential to be powerful learning tools. At this point, most educators are not prepared to deliver curricula that derive the most ethical and educationally powerful benefits from AI content platforms.

1Colorado State University, Pueblo CO 81001, USA

This paper examines 22 studies addressing the efficacy of six primary methods for detecting AI assistance in academic submissions by postsecondary students. Current research suggests that postsecondary students from a range of disciplines use AI-assisted content. Studies also indicate the inability of any one detection method or combination of methods to adequately identify AI submissions. The literature suggests that AI programs have the potential to create high-quality work, reinforcing the value of AI-assisted learning if supported by learning methods. Given that (1) students use and will continue to use AI support, (2) AI-generated submissions are difficult to detect, and (3) quality education and methods are available to support AI-assisted learning, institutions, and educators, there is a critical and immediate need to implement guidelines on AI ethics in course introductions for all curricula. While the following suggestions were formulated with university graduate-level business students in mind, the core concepts could be applied to all disciplines.

What follows are four basic steps for educators and curriculum developers to take to support the ethical and educationally positive student use of AI platforms to complete assignments:

- First, each course should begin with a candid presentation outlining your approach to AI in research and studies. Include the methods that you use and be upfront with your students. Leading state universities have incorporated language into course syllabi to guide students' ethical use of AI for research and studies [7]. This language may be tailored to individual class dynamics. What is most important is to listen to student voices when crafting protocol. We must understand why students use AI: Is it to save time, to get better grades, or to improve their skills? Students' desires to learn and improve their skills should be supported.
- The second step is to maintain up-to-date awareness of all AI platforms and adopt a positive, structured policy for accommodating the use of AI in higher education. The instructor and curriculum developer should also be active users of AI research technology. For my courses, I follow Elicit.org, a non-profit AI research assistant that currently searches Semantic Scholar and provides information on more than 125 million research papers [8]. I offer a live video presentation that covers AI use, protocol, and best practices at the beginning of each course.
- Third, course assignments and discussion questions should be designed in a manner that supports critical thinking by asking the student to show/tell "why" and by asking the student "what if" questions [9]. Educators must initiate candid conversations with students that empower them to be critical thinkers [10]. In addition to beginning each class with a structured AI literacy and ethics segment that covers research techniques, tools for verifying information and methods for preventing plagiarism should be shared. When it comes to AI ethics in teaching and learning, there need to be opportunities for students who may use AI for submissions yet are unfamiliar with the basics of AI technology to access a well-structured AI literacy education [11]. If we—educators—are to best prepare future business leaders to ethically use and understand AI, the curriculum should include discussions and information on human reasoning versus analogical reasoning.

As educators, we are not doing our job if we fail to prepare students for the ethical challenges raised by AI technology in the business environment [12]. Ultimately, human developers will most likely continue to be held accountable for the ethical lapses of AI processes [13]. Business graduates must be prepared to ethically use AI and co-evolve with further AI developments in a way that most benefits both the individual and humankind.

Acknowledgments: The AI research generator Elicit.org was used to search for academic papers in Semantic Scholar. Five hundred papers were initially identified as relevant to this study, which were narrowed down to 22 full-text studies.

Disclosure of Interests: The author has no competing interests to declare that are relevant to the content of this article.

Table 1. Thematic Analysis of Detection Technologies' Effectiveness

Detection Tool Type	Accuracy Range	Limitations	Key Features
AI-specific detectors (e.g., GPTZero, Turnitin AI)	0%–100%	False positives, vulnerability to obfuscation	Specialized algorithms for AI text detection
Plagiarism detection software (e.g., Turnitin, Moss)	Varied	Less effective for AI-generated content	Similarity checking, database comparison
Machine learning models	90%	Require large datasets, potential overfitting	Adaptability, pattern recognition
Human evaluation	35%–69%	Inconsistent, time consuming	Contextual understanding, qualitative assessment
Keyword analysis	Not included	Limited scope, potential false positives	Simplicity, ease of implementation
Multi-tool approaches	79%–91%	Complexity, potential contradictions	Comprehensive coverage, cross-validation

Table 2. Included Studies [14]

Study	Population Studied	Detection Method	Sample Size	Detection Accuracy
Chaka, C., 2024	Undergrad English	30 AI detectors	40 essays	Varied (0%–100%)
Murray, N., & Tersigni, E. 2024	Postsecondary writing	Human evaluation	20 instructors	35% overall
Chen et al., 2024	Introductory programming	Plagiarism markers computational analysis	983 students	85% detection rate
Duane, 2024	Undergraduate business	Turnitin AI detection	21 tests	23.8% passed undetected
Gosling et al., 2024	Undergraduate psychology	Turnitin AI detection	320 responses	94.7% accuracy

Ibrahim et al., 2023	University-wide	GPTZero OpenAI Classifier	1,601 participants	Varied (51%–95%)
Jambunathan et al., 2024	Non-university	ZigZag ResNet	Unknown	88.35%
Karnalim et al., 2023	Web programming	Not specified	16 participants	Unknown
Kost, 2024	Doctoral studies	Turnitin AI detection	48 papers	Unknown
Krecar et al., 2024	Multiple disciplines	Human evaluation	350 students 12 faculty	53.75%
Mapletoft et al., 2024	Undergraduate business	Multiple AI detectors	Not specified	Varied (0%–75%)
Mozelius, 2024	Master's program	Multiple AI detectors	16 students	Varied (46.6%–71.8%)
Paustain and Slinger, 2024	Undergraduate microbiology	Multiple AI detectors	153 students	79%–91%
Peng et al., 2024	High school essays	Multiple AI detectors	Not specified	> 90% for unperturbed text
Perkins et al., 2023	Higher education	Turnitin AI detection	22 submissions	91% flagged, 58% identified
Perkins et al., 2024	Undergraduate	Multiple AI detectors	805 samples	39.5% (unmanipulated) 17.4% (manipulated)
Price and Sakellarios, 2023	Undergraduate	Multiple AI detectors	3 students	Varied (0%–99%)
Scarfe et al., 2024	Undergraduate psychology	Human evaluation	Not specified	6% detection
Streletska et al., 2024	Undergraduate STEM	GPT detector	Not specified	7.46% and 4.03% violation rates
Talaue, 2023	Undergraduate English for academic purposes (EAP)	Turnitin AI detection	5 students	29%–100% AI use
Taylor et al., 2023	Undergraduate computer science	Moss plagiarism detection	59 students, 150 AI samples	Not adequately detected
Want et al., 2023	Computer science	Human evaluation, automated tools	10 evaluators	Human: 69% accuracy, Tools: varied
Weber-Wulff et al., 2023	Various disciplines	14 AI detection tools	Not specified	Below 80%
Weichert and Dimobi, 2024	Undergraduate – varied	Multiple AI detectors	212 human essays	Varied (0%–24.64%)

References

1. Alkhalifah, J.M., Bedaiwi, A.M., Shaikh, N., Seddiq, W., Meo, S. A.: Existential anxiety about artificial intelligence (AI) -is it the end of humanity era or a new chapter in the human revolution: questionnaire-based observational study. Frontiers in Psychiatry 15, 1368122–1368122 (2024) https://doi.org/10.3389/fpsyt.2024.1368122
2. Chao, W.: Yang Qiang, interviewed by Wang Chao. In: UNESCO Currier, The Fourth Revolution (June 2018) https://courier.unesco.org/en/articles/fourth-revolution

THE ROLE OF THE GREAT POWERS IN INTERNATIONAL POLITICS. TRUMP AND ETHICS IN THE INTERNATIONAL ARENA

PhD. Flavio Rafael González Ayala[1] [0000-0001-7317-2485]

Keywords: International Ethics, International Politics, Trump, United States, China, Tariffs.

Abstract

This article aims to present ethics in current international politics, primarily manifested in President Donald Trump's second term in the United States and its repercussions on the world stage.
This analysis will be conducted from the perspective of realist, constructivist, and institutionalist theories of international relations.
The economic impact of the tariffs imposed by the Trump administration on world trade is reviewed. It also examines the threats to incorporate Greenland into the United States and reclaim the Panama Canal, as well as its intervention in the peace agreements between Russia and Ukraine, viewed from the aforementioned theories, incorporating national security concerns and from the perspective of Thucydides' theory. This theory will compare the actions of the United States, China, Russia, and the European Union.
This article presents the ethics of international policy primarily those of Trump and some current powers. However, the recommendations proposed by the various theories reviewed require modifying some stated policies. The aim is to understand these situations and show the path the world is headed in if certain actions, mentioned in the conclusions, are not taken.

Introduction

Donald Trump's second term appears to represent a turning point in US foreign policy, with a focus on ethics behind his decisions. You could also briefly explain the international relations theories that will guide the analysis (realism, constructivism, and institutionalism) and how they apply to Trump's decisions in particular. The question of whether it is ethical for great powers to use their influence to impose their interests on other countries, as in the case of Trump's threatened tariffs on Mexico, Canada, and the world, or economic sanctions and military interventions, is an issue that involves a delicate balance between political realism, morality, and international justice.

From a realist perspective, nations act primarily in their own national interests, seeking to maximize their power and security in an anarchic international system.

1 Universidad Autónoma de San Luis Potosí (UASLP), SLP, México.

According to this view, great powers, such as the United States, have the right (and sometimes the duty) to use their economic, diplomatic, and military power to protect their interests. Within this framework, sanctions, tariffs, or even military interventions can be viewed as legitimate tools to achieve these goals, contrary to constructivist or institutionalist theory of international relations.

Background

Current policies and threats of President Donald Trump's second term

The second administration of United States President Donald Trump began on January 20, 2025. However, since the election results were announced in November 2024, he has launched a series of threats, including imposing tariffs on his trading partners and all other countries, reclaiming the Panama Canal, incorporating Greenland into the United States, and ceasing to support Ukraine in its war, among others.

The justification began with his "America First" policy, which, during his first term from 2017 to 2021, resulted in the withdrawal from international treaties such as the Paris climate agreement and the Iran nuclear deal, trade tensions with China, disagreements with traditional allies such as the North Atlantic Treaty Organization (NATO), and his stance on concessions to Russia.

Another of his controversial actions is his immigration policy, which he considers the United States to be full of enemies who have entered the country illegally.

Ethics versus US policies

Ethics is the science that aims to direct human actions and our will toward goodness, true happiness, the perfection of the individual, and, consequently, the prosperity of societies. (UNAM, 2025). Regarding Trump's international policy, the interests of the United States are prioritized over global concerns, such as the environment, political stability, and world peace. This section will delve deeper into this topic.

Theories of International Relations

Realism, constructivism, and institutionalism. How the authors of these theories express themselves regarding Trump's international policy actions.

Thucydides' Theory

Thucydides' theory is based on the idea that when an emerging power (in this case, China) challenges an established power (the United States), this generates tensions that can lead to conflict. This point is crucial for analyzing Trump's behavior on the international stage.

It is necessary to compare the policies of Trump, China, Russia, and the EU, analyzing whether Trump's behavior reflects a perception of the decline of US power and an attempt to reaffirm its global position.

Tariff Threats

The trade war between the United States and China could be analyzed from several perspectives. The use of tariffs is a clear example of Trump's policy and its focus on protecting the US economy.

The Artificial Intelligence market and competition between the United States and China.

Threats to incorporate Greenland into the United States and to reclaim the Panama Canal.

Russia-Ukraine War and US Mediation. What does the United States intend with this mediation?

The ethics of the United States, China, Russia, and the European Union in their participation on the international stage.

Conclusions

The article concludes that Trump's policies, especially in his second term, reflect an approach focused on US national interests, often to the detriment of global cooperation and international stability. Recommendations based on the theories analyzed suggest that a more ethical and cooperative approach is needed to avoid further global conflict and ensure long-term global prosperity.

References

1. Waltz, Kenneth N. (1979). Theory of International Politics. Addison-Wesley.
2. Morgenthau, Hans J. (1948). Politics Among Nations: The Struggle for Power and Peace. Alfred A. Knopf.
3. Wendt, Alexander. (1999). Social Theory of International Politics. Cambridge University Press.
4. Ruggie, John Gerard. (1998). Constructing the World Polity: Essays on International
5. Finnemore, Martha. (1996). National Interests in International Society. Cornell University Press.
6. Keohane, Robert O. (1984). After Hegemony: Cooperation and Discord in the World Political Economy. Princeton University Press.
7. Allison, Graham. (2017). Destined for War: Can America and China Escape Thucydides's Trap?. Houghton Mifflin Harcourt.
8. Trump, Donald J. (2015). Crippled America: How to Make America Great Again. Threshold Editions.
9. Sanger, David E. (2020). The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age. Crown Publishing Group.
10. Bacevich, Andrew J. (2020). The Age of Illusions: How America Squandered Its Cold War Victory. Metropolitan Books.
11. Zakaria, Fareed. (2019). Ten Lessons for a Post-Pandemic World. W.W. Norton & Company.
12. Brooks, Stephen G., & Wohlforth, William C. (2016). The Trump Doctrine: The Foreign Policy of the "America First" President. Foreign Affairs.
13. Baldwin, Robert E. (1985). The Political Economy of US Tariffs. MIT Press.
14. Ghosn, Fadil, & Wilson, David. (2020). Global Trade and the US-China Economic Rivalry. Routledge.
15. Chanda, Nayan. (2020). The China-United States Trade War and Future Economic Relations. Routledge

16. Mele, Christopher. (2019). Trump's Greenland Idea and the History of U.S. Interest in the Arctic. Journal of Arctic Studies.
17. Strain, Jennifer. (2018). The U.S. and the Panama Canal: A Historical Relationship. Pan-American Review.
18. Mearsheimer, John J. (2014). Why the Ukraine Crisis Is the West's Fault: The Liberal Delusions That Provoked Putin. Foreign Affairs.
19. Rice, Condoleezza. (2011). No Higher Honor: A Memoir of My Years in Washington. Crown Publishers.
20. Rumer, Eugene, & Sokolsky, Richard. (2015). U.S.-Russian Relations: From Cold War to Cold Peace. Carnegie Endowment for International Peace.
21. Rawls, John. (1999). The Law of Peoples. Harvard University Press.
22. Nardin, Terry. (2005). The Ethics of International Relations. University of California Press.
23. Bessner, Daniel. (2019). The Ethical Dimension of U.S. Foreign Policy. International Security Journal.
24. Fitzgerald, W. J. (2018). The Ethics of Intervention: Global Responsibility and U.S. Foreign Policy. Global Policy Journal.
25. Kaplan, Robert D. (2019). The Return of Marco Polo's World: War, Strategy, and American Interests in the Twenty-first Century. Random House.

THE MEDIATING EFFECT OF JOB CRAFTING IN THE RELATIONSHIP BETWEEN ORGANIZATIONAL COMMITMENT AND ORGANIZATIONAL CITIZENSHIP BEHAVIOR AND ITS ETHICAL IMPLICATIONS

Melissa Eugenia Treviño Serna[1] [0009-0003-7629-5108],
Rosalba Martínez Hernández[2] [0000-0003-3721-6094],
Juan Carlos Yáñez Luna[3] [0000-0001-8341-9015]

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

Extended Abstract

The study of job crafting and his 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 been more proactive in their duties [1]; the increasing interest in the use of creativity by employees as a mean 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.

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.

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

1 Universidad Autónoma de San Luis Potosí, San Luis Potosí 78000, México, melissa.trevino@uaslp.mx
2 Universidad Autónoma de San Luis Potosí, San Luis Potosí 78000, México
3 Universidad Autónoma de San Luis Potosí, San Luis Potosí 78000, México

can lead to positive outcomes in organizations such as: profits, retention rates and better customer service [12].

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 [13].

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 [14].

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 [15].

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

Considering this, Abuzaid & Ghadi [16] 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 [17] 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.

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 [18] due to it centers on the analysis of total variance and focuses on measuring the model and its structure. Based on PLS-SEM the following relationships can be tested as:

1. The direct effect of organizational commitment on job crafting.
   a. Hypothesis: Organizational Commitment has a direct effect on job crafting.
2. The direct effect of job crafting on organizational citizenship behavior.
   a. Hypothesis: Job Crafting has a direct on effect Organizational Citizenship Behavior.

1. The direct effect of organizational commitment on organizational citizenship behavior.
   a. Hypothesis: Organizational Commitment has a direct on effect Organizational Citizenship Behavior.
2. The mediating effect of job crafting in the relationship between organizational commitment and organizational citizenship behavior.
   a. Hypothesis: Job Crafting has a mediating effect on the relationship of Organizational Commitment and Organizational Citizenship Behavior.

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.

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

References

1. Schachler, V., Epple, 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.
2. Tian, W., Wang, H., Rispens, 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.
3. Bala, S., 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, 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 Manag. 26, 1–24 (2016). https://doi.org/http://dx.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.: Digital Transformation in the Hospitality Industry : Improving Efficiency and Guest Experience. Int. J. Manag. Sci. Inf. Technol. 4, 428–437 (2024). https://doi.org/https://doi.org/10.35870/ijmsit.v4i2.3201.
12. Gallup Consulting: State of the Global Workplace,Washington, D.C. (2024).
13. DNV & United Nations Global Compact: IMPACT Transforming business, changing the world. Høvik, Norway (2024).
14. 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.
15. OECD: OECD Guidelines for Multinational Enterprises on Responsible Business Conduct.OECD Publishing, Paris (2023). https://doi.org/10.1787/81f92357-en.
16. 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. Adm. Sci. 14, 1–20 (2024). https://doi.org/. https://doi.org/10.3390/admsci14090191.
17. Consejo Coordinador Empresarial (CCE): Código de Integridad y Ética Empresarial. México (2017).

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Black, W.C., Anderson, R.E.: Multivariate Data Analysis. 8th edn.Cengage Learning EMEA,Hampshire, UK (2018). https://doi.org/10.1002/9781119409137.ch4.

CORPORATE GOVERNANCE, BUSINESS ETHICS, AND THEIR RELATIONSHIP WITH FINANCIAL PERFORMANCE IN COMPANIES LISTED ON THE MEXICAN STOCK EXCHANGE

Francisco Eduardo Pérez Ortega[1], Pedro Isidoro González Ramírez[2] [0000-0002-0763-954X], Guadalupe del Carmen Briano Turrent[3] [0000-0001-8241-0385], Juan Carlos Yáñez Luna[4] [0000-0001-8341-9015]

Keywords: Corporate Governance, Business Ethics, Financial Performance, Mexican Stock Exchange.

Introduction

The beginning of the 21st century has been marked by a series of significant changes in business practices globally. Increased competitiveness, driven by globalization, has led companies to adopt strategic approaches aimed at generating long-term value (Mukhtaruddin et al., 2019). However, corporate fraud scandals such as those involving Enron, Xerox, and Parmalat have undermined investor confidence, prompting the implementation of regulatory frameworks like the Sarbanes-Oxley Act in the United States to prevent financial irregularities.

In this context, business ethics has become an essential pillar for organizational sustainability. Ethical decisions not only have repercussions on reputation and investor confidence but are also directly linked to long-term financial performance. Companies that adopt ethical practices in their corporate governance tend to foster an environment of transparency, accountability, and trust, improving their market positioning and, in turn, their financial outcomes. This ethical approach benefits not only shareholders but also all stakeholders, including employees, customers, and society at large.

In the Mexican context, business transparency still faces significant challenges due to the lack of mandatory regulations for the disclosure of ESG (Environmental, Social, and Governance) information. The absence of standardized ESG metrics affects investor confidence (IFAC, 2023). Additionally, structural issues such as corruption, high concentration of corporate ownership, and limited employee representation on boards of directors have hindered the development of stock markets and impacted the financial performance of companies (Watkins Fassler C Flores Vargas, 2015).

Given these challenges, the role of corporate governance becomes crucial in promoting and ensuring transparency and accountability in business practices. Regulatory

1 Facultad de Economía, Universidad Autónoma de San Luis Potosí. San Luis Potosí, México, a281034@alumnos.uaslp.mx
2 Facultad de Economía, Universidad Autónoma de San Luis Potosí. San Luis Potosí, México, pedro.gonzalez@uaslp.mx.
3 Facultad de Contaduría y Administración, Universidad Autónoma de San Luis Potosí., San Luis Potosí, México, guadalupe.briano@uaslp.mx
4 Facultad de Economía, Universidad Autónoma de San Luis Potosí. San Luis Potosí, México, jcyl@uaslp.mx

frameworks guarantee that activities are conducted ethically, while corporate governance plays a key role in protecting the interests of both minority and majority shareholders. One of its primary objectives is to ensure that companies are managed efficiently and effectively to maximize profits in a sustainable manner (Vyas, 2023).

The primary goal of this study is to conduct an empirical analysis to evaluate the relationship between governance dimensions and their impact on the financial performance of companies in the non- financial sectors listed on the Mexican Stock Exchange (BMV) between 2017 and 2022. This will involve evaluating the impact of governance both individually and collectively through descriptive statistical analysis and multiple regression analysis. Based on agency theory, one of the main theories in finance and corporate governance (Tekin C Polat, 2020), and stakeholder theory, which posits a positive relationship between corporate governance and financial performance (Schäfer, 2019), the general hypothesis is that the governance dimension of ESG indicators positively influences the profitability of these companies.

Research Objectives

1. To construct a corporate governance index based on Mexican regulations and international literature.
2. To evaluate the impact of each element of corporate governance on the financial performance of non-financial companies listed on the BMV.
3. To assess how different levels of corporate governance compliance impact key financial metrics, such as profitability.
4. To propose strategic recommendations for improving corporate governance practices based on the study's findings.

In the context of an increasingly digitalized business environment, the ethical challenges related to corporate governance take on new dimensions. Digitalization not only transforms operational processes but also influences how companies implement and communicate their ethical and governance policies. Although this study focuses on non-financial companies in Mexico, its findings have relevant implications for the digital world, particularly regarding transparency, the traceability of decision-making, and the integrity of information shared through digital platforms. Therefore, this analysis also contributes to the broader discussion on ethical challenges for business in a digital world, offering empirical insights into how governance practices can be adapted and strengthened in the information age.

Methodology

A quantitative approach based on econometric models and a longitudinal design is employed, analyzing data from 101 non-financial companies listed on the BMV between 2017 and 2022. The database was constructed manually from annual, and sustainability reports available on the BMV and official company websites.

meetings, and promote the inclusion of women on boards of directors. Additionally, it is suggested to review and adjust minority shareholder protection policies and codes of conduct to minimize negative effects, particularly in areas such as liquidity and profitability, ensuring decisions favor a balance between the interests of all shareholders.

References

1. International Federation of Accountants. (2023). Principios revisados de gobierno corporativo de la OCDE: ¿Qué pueden hacer las empresas en América?. https://www.ifac.org/knowledge- gateway/discussion/principios-revisados-de-gobierno corporativo-de-la-ocde- que-pueden- hacer-las-empresas-en-america
2. Mukhtaruddin, M., Ubaidillah, U., Dewi, K., Hakiki, A., C Nopriyanto, N. (2019). Good corporate governance, corporate social responsibility, firm value, and financial performance as moderating variable. Indonesian journal of sustainability accounting and management, 3(1),5, https://doi.org/10.28992/ijsam.v3i1.74
3. Schäfer, H. (2019). Understanding how stakeholders are affecting sustainability and finance. In: on values in finance and ethics. Springerbriefs in finance. Springer, cham. https://doi.org/10.1007/978-3-030-04684-2_5.
4. Tekin, H., C Polat, A. Y. (2020). Agency theory: a review in finance. Anemon muş alparslan üniversitesi sosyal bilimler dergisi, 8(4), 1323–1329. https://doi.org/10.18506/anemon.712351
5. Vyas, S. (2023). Corporate governance practices in India: problems C importance. In ijfmr23032801 (vol. 5, issue 3). www.ijfmr.com
6. Watkins Fassler, K., C Flores Vargas, D. R. (2016). Determinantes de la concentración de la propiedad empresarial en méxico. Contaduria y administracion, 61(2), 224–242. Https://doi.org/10.1016/j.cya.2015.05.015

ETHICAL AND TECHNOLOGICAL PERSPECTIVES ON RISK DISTRIBUTION IN CRYPTOCURRENCIES

Oliver René Arroyo Leos[1] [0000-0002-7448-6328], Cuauhtémoc Modesto López[2] [0000-0002-1906-7219], José Luis de la Fuente García[3] [0009-0007-4982-0797]

Keywords: VIX, Bitcoin, Ética

Abstract

In the stock market there is a Volatility Index better known as VIX, it is a financial indicator developed by the Chicago Board Options Exchange (CBOE), which measures the short-term volatility of the stock market based on the S&P 500 index, this index reflects the perception that investors have about uncertainty or volatility in the stock market. On the other hand, Bitcoin (BTC) is a decentralized cryptocurrency and the first to be created and formally introduced to the market, BTC is the leading cryptocurrency, in terms of market capitalization, whose operation is fundamentally based on the use of Blockchain technology, this financial instrument characterized by its high volatility, given that its price fluctuates significantly in short periods of time, has become an attractive investment for an important group of investors. This research analyzes the econometric relationship between the VIX, as an independent variable and the price of BTC as a dependent variable, in a time horizon ranging from April 20, 2011 to January 31, 2025, with the main objective of observing if there is a negative correlation between the VIX and BTC, demonstrating that an increase in the first variable corresponds to a decrease in the second, with the objective of demonstrating that in the absence of ethical regulations and security mechanisms in the use of BTC, investors in this type of instruments will not care about the volatility that exists in the world, because the BTC becomes a refuge for bad financial practices, making it necessary to implement regulatory, ethical and transparent frameworks for its use, especially in contexts of public investments or business strategies.

Introducción

In the last two decades, financial markets have undergone a series of significant transformations driven by technological advances, globalization, and, above all, the

1 Universidad Autónoma de San Luis Potosí, Mex., oliver.arroyo@eco.uaslp.mx
2 Universidad Autónoma de San Luis Potosí, Mex., cuauhtemoc.modesto@uaslp.mx
3 Universidad Autónoma de San Luis Potosí, Mex., joseluis.delafuente@uaslp.mx

emergence of financial instruments. Among these new instruments, the so-called cryptocurrencies stand out, referring to decentralized digital assets that have gained popularity due to their great potential to offer various alternatives to traditional financial systems. Among them, the one that has stood out the most is Bitcoin (BTC), a cryptocurrency that has become the leader according to market capitalization. This has caused investors, retailers, and financial institutions to seek them as a refuge during periods of economic uncertainty [3], so their volatile behavior causes the search for new investment strategies and, above all, new financial market regulation.

On the other hand, the volatility index (VIX), developed by the Chicago Board Option Exchange (CBOE) [2], is a recognized indicator to determine the perception of risk in traditional financial markets, this is because this index measures the implicit will of the options of the S&P 500 index, which is why it is commonly used as a gauge market uncertainty in real time. For the purposes of this research, the horizon is a period of time that goes from April 20, 2011, to January 31, 2025 [3], even though the BT formally entered the market in 2008, it was not until April 2011 when it began to register growth in a potential way [4], whose relationship with the VIX has not yet been fully studied, much less understood. This space in the academic literature motivates the present research, seeking to analyze the econometric relationship between the VIX and the price of BTC.

The central problem of this research lies in the need to understand how the perception of uncertainty in traditional financial markets, measured by the VIX, influences the behavior of BTC, an asset characterized by its high volatility and its role as a speculative safe haven. This relationship not only has theoretical implications for the development of robust predictive models, but also practical ones for the formulation of regulatory and ethical policies that mitigate the risks associated with the use of BTC in financial transactions. Furthermore, given the growing adoption of cryptocurrencies by individuals, small and medium-sized businesses, and even public entities, it is crucial to ensure the transparency and traceability of these transactions to protect less informed participants and reduce exposure to systemic risk in domestic markets.

The central hypothesis of this research is that there is an inverse (negative) relationship between the VIX and the price of BTC, which would imply that an increase in the perception of uncertainty in traditional markets would be associated with a decrease in the price of BTC. This is based on studies carried out by Bouri et al. (2023), which mention the negative relationship between traditional assets and cryptocurrencies during periods of financial crisis. However, because BTC is one of the newest financial instruments, whose behavior has not been fully studied, it is expected that this study will provide empirical evidence that will help in understanding its dynamics.

If the fact that there is an inverse relationship between the VIX and the BTC is true, it could be thought that given an increase in the perception of risk by investors, they would seek refuge in non-traditional financial instruments such as the BTC with all its implications, since Ferrero and Hoffman (2019) define ethics in the financial field as a set of moral principles and norms that guide the behavior of individuals and organizations among financial markets, which is why the existence of an ethical culture that guarantees the integrity of the financial system is necessary [7].

So much so that since that distant 2008, the perception of the investing public about the morality that lies in the markets has been suffering a deterioration, which

implies an important and urgent need to define ethical standards (Ferrero & Hoffman, 2019), and the financial institutions themselves must act as investment promoters with responsible conduct with the companies in which the investments are carried out (Jackson, 2020; Arjona & De Frutos, 2023), with which the confidence of the participants in the stock market would increase and in turn reduce the volatility that is associated with speculative practices that are not responsible, thereby generating positive externalities for economic and financial development [1], given the above, Sánchez-Pérez and Vargas-Herrera (2021) comment that supervision by government institutions favors the emergence of unethical operations, which generates inefficiencies eroding the credibility of the financial system.

For this reason, it is necessary to make visible the need for regulatory entities and ethical mechanisms that contribute, on the one hand, to providing greater certainty to the operations of this instrument and, on the other, clarify the way in which it operates, in order to reduce the risk exposure of participants, especially those who allocate part of their assets to these operations, or as in the case of small and medium-sized companies that can base their operating and growth strategies on these [4].

Methodology and results

This research has a quantitative approach, which was based on empirical data used to carry out, first of all, a simple review, with the initial objective of evaluating the relationship that exists between the variable BTC as a dependent variable, and Vix as an independent variable, these data come from a sample of 3,500 observations, using the statistical model in general form:

$$y = \beta_0 + \beta_1 x_1$$

Where y is the dependent variable (BTC), β_0 represents the intercept, while x_1 is the independent variable (VIX), and finally β_1 refers to the slope of the model.

In order to better observe the variables and facilitate the interpretation of their relationships and because the variables presented asymmetric distributions, a transformation of the daily growth rates of each one to a natural logarithmic form was carried out in order to reduce the impact of extreme values and try to normalize the distribution of the data. The model was carried out through the least squares method (OLS), with the following results: the coefficient of determination R2 yields a value of 0.0054 which implies that the model only explains 0.54% of the total variability of the BTC, being that the statistic F with a value of 19.128, and an associated p-value that does not suggest an overall significance of the model, so it can be identified that there is a weak relationship between the variables analyzed.

Given the above, and regarding the OLS, it can be observed that, although there is a statistically significant relationship, it is very weak between the VIX and the BTC, where the multiple termination coefficient (R) has a value of 0.0737 R2 of 0.0054, as observed in the previous paragraph, so the model suggests that the explanatory power of the model is very limited and that there are variables that were not includ-

ed in the analysis that could have a greater weight in explaining BTC. In any case, the standard error of the model reached 0.0804, which reflects a high dispersion of the residues around the estimated regression line, within the significance test of the model (ANOVA), the F statistic presents a value of 19.13 with a p-value less than 0.05, which demonstrates a level of statistical significance, however, and due to the R2 result, the result should be taken with caution. The general results of the OLS model are shown in the following table:

Table 1. Results of the application of the MCO with the studied variables

Regression statistics	
Multiple correlation coefficient	0.073746017
Coefficient of determination R^2	0.005438475
Adjusted R^2	0.005154152
Typical error	0.080382297
Observations	3500

Table 2. Results of the application of the MCO with the studied variables continued

	Coefficients	Typica Error
Interception	0.005629247	0.001359685
Variable x_1	-0.071721672	0.016399015

Due to the result obtained with the MCO, a multiple linear regression model was carried out with an AR model which shows that there is statistical significance and also with a very reduced explanatory power like the previous model, since the correlation coefficient (R) reached a value of 0.0839, while the coefficient of determination R2 was 0.0070 which is an indication that only 70% of the variability in BTC is explained by the independent variables of the model (x1 y x2).

Like the previous model, the value of R2 may indicate that the model fit is limited and that there are factors that were not considered in the analysis and that could have a significant influence on the BTC, even though the F statistic of 13.0694 and with a p-value of 0.000002 shows that it is statistically significant on this last variable.

The intercept of the model was estimated at 0.00677 which is statistically significant at the 1% level, where it is explained that when both independent variables take the value of zero, the value of "y" is positive and different from zero, for x1 the result is negative, with a significance at the 5% level and an estimated value of 0.0422 with a p-value of 0.0102, indicating that an increase of 1 in x1. It brings as a consequence, an average decrease of 0.0422 units of "y" as long as x2 remains constant, on the other hand, this variable also reflected a negative and significant effect on "y" with a coefficient of -0.0762 and a p-value of 0.0000087 reflecting that increases in x2 are related to decreases in the values of the variable "y". In tables 3 and 4, the statistics produced by the model are presented.

Table 3. Results of the application of AR with the variables studied

Regression statistics	
Multiple correlation coefficient	0.083902095
Coefficient of determination R^2	0.007039562
Adjusted R^2	0.006500934
Typical error	0.085106435
Observations	3690

Tabla 4. Results of the application of AR with the variables studied

	Coefficients	Typica Error
Interception	0.006766292	0.00140579
Variable x_1	-0.042191817	0.016412596
Variable x_2	-0.076206977	0.017107736

Even though both coefficients are statistically significant, their lower magnitude indicates that the impact they have on "y" is minimal, so their usefulness in predicting the model in empirical applications is limited.

Conclusions

The research was carried out in relation to the Stock Market Volatility Index (VIX) and the price of Bitcoin (BTC), considered an emerging financial asset with certain characteristics of speculative refuge, for this purpose two regression models were implemented, with 3,500 data in a research horizon of April 2011 and January 2025, the first model was with least squares (OLS) and the second with an autoregressive model (AR).

The results obtained show that there is an inverse or negative relationship between the variables and although it is statistically significant, the relationship also shows a weak relationship, and although the explanatory power of the models was limited, with coefficients of determination R2 less than 1%, which in some way supports the initial hypothesis, the initial hypothesis is partially supported.

From an ethical and regulatory perspective, the findings demonstrate the importance of strengthening regulatory frameworks that incorporate sound ethical principles in the international financial market, particularly with regard to the adoption of cryptocurrencies, since these financial assets can be a safe haven for malpractice and money laundering.

Future lines of research could include incorporating additional relevant variables, applying nonlinear models and advanced machine learning techniques, and examining the regulatory impact and ethical framework on the stability of the cryptocurrency market, with the aim of observing the relationship this market may have with the ethics of financial markets.

References

[1] R. Al-Adeeb, A. Ansong, Ethical investing and financial markets: A systematic literature review. J. Bus. Ethics 173(2), 265–284 (2021). https://doi.org/10.1007/s10551-021-04786-z

[2] M. Arjona, C. de Frutos, Sustainable finance and ethical investment: Regulatory challenges and opportunities. Eur. Bus. Organ. Law Rev. 24(1), 123–147 (2023). https://doi.org/10.1017/S156675292300012X

[3] E. Bouri, R. Gupta, A.K. Tiwari, Y. Wang, The relationship between traditional assets and cryptocurrencies during financial crises: Evidence from advanced econometric models. J. Financ. Mark. 45(3), 100–115 (2023). https://doi.org/10.1016/j.finmar.2023.100115

[4] CEPAL, El papel de las criptomonedas en las estrategias de crecimiento de pequeñas y medianas empresas en América Latina. Comisión Económica para América Latina y el Caribe (2021). https://www.cepal.org

[5] I. Ferrero, W.M. Hoffman, Financial ethics in the post-crisis era: Trust, integrity and market behavior. Bus. Soc. Rev. 124(4), 549–570 (2019). https://doi.org/10.1111/basr.12183

[6] G. Jackson, Regulating ethical behavior in financial markets: The role of institutional investors. Corp. Gov. Int. Rev. 28(5), 334–347 (2020). https://doi.org/10.1111/corg.12321

[7] J. Martínez-Ferrero, I.-M. García-Sánchez, The impact of ethical culture on market integrity and investor confidence. Financ. Res. Lett. 46, 102489 (2022). https://doi.org/10.1016/j.frl.2022.102489

[8] M. Sánchez-Pérez, J. Vargas-Herrera, Insider trading and ethical dilemmas in emerging capital markets: Evidence from Latin America. Emerg. Mark. Rev. 49, 100836 (2021). https://doi.org/10.1016/j.emr.2021.100836

[9] A. Zohar, H. Dinçer, Algorithmic trading and ethical risks in modern financial markets. Technol. Forecast. Soc. Change 199, 123098 (2024). https://doi.org/10.1016/j.techfore.2024.123098

GEOPOLITICAL AND ETHICAL PERSPECTIVES OF CLOUD COMPUTING

Ala Ali Almahameed[1] [0000-0002-4381-0316], Jorge Pelegrín-Borondo[2] [0000-0003-2720-1788], Juan Luís López-Galiacho Perona[3] [0000-0001-9364-3539], Mario Arias-Oliva[4] [0000-0002-6874-4036]

Keywords: Cloud Services, Security, Privacy, Data Sovereignty

Extended Abstract

Cloud computing has become a cornerstone of modern business operations, offering organizations across the globe unparalleled benefits such as scalability, cost efficiency, and enhanced collaboration. However, its adoption is not without significant challenges, which vary by region due to differences in regulatory environments, infrastructure, and economic conditions. In the United States, cloud computing has been widely adopted across industries, driven by its ability to reduce operational costs, improve flexibility, and foster innovation. However, its implementation is accompanied by several challenges that organizations must navigate to fully realize its potential. One of the foremost challenges in the U.S. is ensuring data security and privacy. The multi-tenant nature of cloud environments raises concerns about unauthorized access, data breaches, and insider threats (Asiri, 2024). Organizations must implement robust security measures, such as encryption and multi-factor authentication, to protect sensitive information. Additionally, the rise of sophisticated cyberattacks, including distributed denial-of-service (DDoS) attacks, underscores the need for advanced threat detection and mitigation strategies (Alghofaili et al., 2021). Compliance with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the Sarbanes-Oxley Act adds complexity to cloud adoption. These regulations require organizations to ensure that their cloud providers adhere to strict data protection standards, which can be challenging given the dynamic nature of cloud environments (Asiri, 2024). Failure to comply can result in legal repercussions and reputational damage.

While cloud computing is often touted for its cost efficiency, many organizations struggle with unexpected expenses due to inaccurate forecasting and ineffective optimization strategies. A Gartner (2024) survey revealed that 69% of IT leaders experienced budget overruns in 2023, highlighting the need for better cost management practices. Organizations must adopt tools and strategies to monitor and optimize cloud usage to avoid financial strain. Effective management of cloud resources remains a sig-

1 University Rovira i Virgili, Tarragona, Spain
2 University of La Rioja, Logroño, Spain
3 Rey Juan Carlos University, Madrid, Spain
4 Complutense University of Madrid, Madrid, Spain

nificant challenge, particularly in hybrid cloud systems. Studies have identified synchronization and backup issues as major technical barriers, which can disrupt operations and reduce efficiency (Hui et al., 2023). Organizations must develop strategies to ensure seamless integration and management of cloud resources. Trust is a critical factor in cloud adoption, particularly regarding data privacy and regulatory compliance. The lack of clear service-level agreements (SLAs) and transparency in cloud service providers' practices can deter organizations from fully embracing cloud solutions (Khan & Ullah, 2016). Building trust requires providers to demonstrate a commitment to security, compliance, and customer satisfaction.

In Europe, cloud adoption is growing, with 43% of EU enterprises having adopted cloud computing services in 2023 (Bauer, Erixon, & Pandya, 2024). However, the region faces unique challenges that stem from its regulatory environment, market dynamics, and infrastructure gaps. Data sovereignty is a primary concern in Europe, driven by stringent regulations such as the General Data Protection Regulation (GDPR). European companies are increasingly turning to sovereign clouds to address privacy issues and ensure compliance with local laws (Telecom Review Europe, 2024). Sovereign clouds allow organizations to store and process data within the EU, reducing the risk of unauthorized access by foreign entities. The dominance of US-based cloud providers, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud, poses significant challenges for European data sovereignty and competition. This dominance has led to concerns about the stifling of European cloud providers and the need for greater digital independence (Bartram, 2024). Policymakers are advocating for reduced reliance on foreign cloud services to promote the growth of native European providers. The European regulatory landscape is complex, with initiatives such as the Digital Services Act (DSA) and Digital Markets Act (DMA) adding layers of compliance requirements for cloud providers. These regulations aim to ensure fair competition and protect user data but can create challenges for organizations navigating the legal framework (Usercentrics, 2024). Security remains a significant concern in Europe, with initiatives like the EU Cloud Code of Conduct aiming to enhance trust among users. The code provides a framework for cloud service providers to demonstrate compliance with GDPR standards, thereby addressing privacy and security concerns (Blancato, 2024). Europe faces a substantial investment gap in ICT and cloud-related sectors compared to the US, totaling approximately USD 1.36 trillion. This gap presents a significant challenge for native European cloud providers to compete effectively with their US counterparts (Bauer, Erixon, & Pandya, 2024). Addressing this gap requires increased investment in infrastructure and innovation.

In the Middle East, cloud adoption is hindered by a combination of security concerns, regulatory challenges, infrastructure limitations, and financial constraints. Despite these barriers, governments in the region are actively promoting digital transformation to foster economic growth. Security and privacy remain paramount concerns for organizations in the Middle East, particularly among small and medium-sized enterprises (SMEs). Apprehensions about data breaches and unauthorized access significantly impede cloud adoption (Alrababah, 2022). Addressing these concerns requires robust security measures and increased awareness of best practices. The absence of comprehensive laws governing technology use creates uncertainties around compliance and data

sovereignty, discouraging businesses from transitioning to cloud platforms (Alrababah, 2022). Governments in the region must develop clear and consistent regulatory frameworks to support cloud adoption. Limited access to cloud computing environments and colocation options hinders the ability of providers to offer a broader range of services. Despite governmental efforts to foster a digital economy, infrastructure gaps remain a significant barrier to cloud adoption (Abudaqqa, 2025). Budgetary restrictions and concerns about return on investment deter SMEs from adopting cloud-based solutions. Additionally, currency fluctuations in the region introduce financial uncertainty, as the cost of cloud services can vary significantly due to exchange rate movements (Arora et al., 2024). Strategies such as local currency contracts or regular price realignment are needed to mitigate this issue.

Data sovereignty refers to the concept that information is subject to the laws and governance structures of the nation where it is collected or processed. As cloud computing enables data storage and processing across global infrastructures, countries have implemented various policies to maintain control over their data. The US has taken a proactive approach to cloud sovereignty through legislative measures such as the Clarifying Lawful Overseas Use of Data (CLOUD) Act, enacted in 2018. This act allows US-based technology companies to provide federal law enforcement access to data stored on servers, regardless of location. While the law facilitates cross-border cooperation in criminal investigations, it has raised concerns among other countries regarding conflicts with their domestic data protection laws (Qin, 2025). The EU has adopted a strong stance on data protection and privacy through GDPR, which enforces stringent rules on data storage, processing, and transfer. The EU Cloud Code of Conduct, introduced in 2021, provides a framework for cloud service providers to ensure compliance with GDPR standards (Blancato, 2024). These efforts reflect the EU's commitment to safeguarding personal data and maintaining trust in cloud computing services.

China has implemented some of the strictest data localization laws in the world, requiring certain categories of data to be stored within its national borders. Under China's Cybersecurity Law, companies handling critical information infrastructure must store data domestically and undergo security assessments before transferring data abroad (Yun, 2024). This approach reflects China's broader strategy of digital sovereignty, where state control over information is prioritized. A survey conducted across multiple countries, including France, Germany, the United Kingdom, and the United States, revealed that 98% of organizations have implemented or plan to implement data sovereignty strategies (Scality, 2023). Many companies are adopting hybrid cloud solutions to store sensitive data within their home country while leveraging global cloud services for other operations. Some organizations are also turning to regional cloud service providers as alternatives to major US-based providers (Malik et al., 2024).

Despite these efforts, significant challenges remain due to the lack of global standardization in data sovereignty policies. The fragmented regulatory landscape creates compliance risks and enforcement difficulties for multinational organizations operating in multiple jurisdictions. Addressing these challenges requires increased international collaboration and the development of harmonized frameworks that balance national security concerns with the benefits of a globally connected digital economy (Lohmöller et al., 2024). As cloud computing continues to evolve, governments world

wide are actively refining policies to assert data sovereignty. The rapid advancement of technology necessitates ongoing discussions between policymakers, industry leaders, and international organizations to establish best practices for data governance. While data sovereignty measures are essential for protecting national interests, achieving a balance between security, privacy, and innovation will be crucial for the future of cloud computing in a digitalized world (Belli, Gaspar, & Jaswant, 2024).

Cloud computing offers immense potential for organizations worldwide, but its adoption is fraught with challenges that vary by region. Addressing these challenges requires a combination of robust security practices, regulatory adherence, efficient resource management, and transparent provider-client relationships. As the global digital economy continues to grow, the importance of data sovereignty and international collaboration will only increase, shaping the future of cloud computing for years to come. In the extended paper, we will discuss further the theoretical framework of geopolitical and ethical perspectives of cloud computing.

References

1. Alghofaili, Y., Albattah, A., Alrajeh, N., Rassam, M. A., & Al-rimy, B. A. S.: Secure cloud infrastructure: A survey on issues, current solutions, and open challenges. Applied Sciences, 11(19), 9005 (2021) https://doi.org/10.3390/app11199005
2. Alrababah, Z.: Barriers to cloud computing adoption among SMEs in the Middle East: A systematic review. Journal of Theoretical and Applied Information Technology, 101(17), 4567-4578 (2022)
3. Asiri, N.: Security and privacy issues in cloud storage. arXiv preprint, arXiv:2401.04076 (2024) https://doi.org/10.48550/arXiv.2401.04076
4. Bauer, M., Erixon, F., & Pandya, D.: The EU gap in ICT and cloud computing. European Centre for International Political Economy (ECIPE). May 2024, from https://ecipe.org/publications/eu-gap-ict-and-cloud-computing/, last accessed 15/02/2025
5. Belli, L., Gaspar, W. B., & Jaswant, S. S.: Data sovereignty and data transfers as fundamental elements of digital transformation: Lessons from the BRICS countries. Computer Law & Security Review, 54, 106017. https://doi.org/10.1016/j.clsr.2024.106017
6. Blancato, F. G. (2024). The cloud sovereignty nexus: How the European Union seeks to reverse strategic dependencies in its digital ecosystem. Policy & Internet, 16(1), 12-32 (2024) https://doi.org/10.1002/poi3.358
7. Hummel, P., Braun, M., Tretter, M., & Dabrock, P.: Data sovereignty: A review. Big Data & Society, 8(1), 2053951720982012 (2021) https://doi.org/10.1177/2053951720982012
8. Kopáčková, H, Htoo, S.T.: Cloud computing services–emerging trends during the times of pandemic. In: 18th Iberian CISTI, pp. 1-6. IEEE, Portugal (2023)
9. Lohmöller, J., Pennekamp, J., Matzutt, R., Schneider, C. V., Vlad, E., Trautwein, C., & Wehrle, K.: The unresolved need for dependable guarantees on security, sovereignty, and trust in data ecosystems. Data & Knowledge Engineering, 151, 102301 (2024) https://doi.org/10.1016/j.datak.2024.102301
10. Malik, A. W., Bhatti, D. S., Park, T. J., Ishtiaq, H. U., Ryou, J. C., & Kim, K. I.: Cloud digital forensics: Beyond tools, techniques, and challenges. Sensors, 24(2), 433 (2024) https://doi.org/10.3390/s24020433
11. Qin, H.: Regulatory Conflict and the Struggle for Digital Sovereignty: A Critical Analysis of the EU-US Data Privacy Framework. Studies in Law and Justice, 4(1), 46-59 (2025) https://doi.org/10.56397/JARE.2023.07.02
12. Greene, J.: New Research on Data Sovereignty. Scality, 13 September 2022, https://www.scality.com/press-releases/new-research-on-data-sovereignty/, last accessed 15/02/2025
13. Yun, H.: China's data sovereignty and security: Implications for global digital borders and governance. Chinese Political Science Review, 1-26 (2024) https://doi.org/10.1007/s41111-024-00269-9

THE USE OF SOCIAL MEDIA AND ARTIFICIAL INTELLIGENCE TO RADICALIZE YOUNG PEOPLE IN JIHADIST TERRORISM

Paula M. Núñez-Guerra[1] [0000-0001-8245-3772]

Keywords: AI, social media, jihadist terrorism, radicalization.

Extended Abstract

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 11M 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 (2024) 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 (2004), 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.

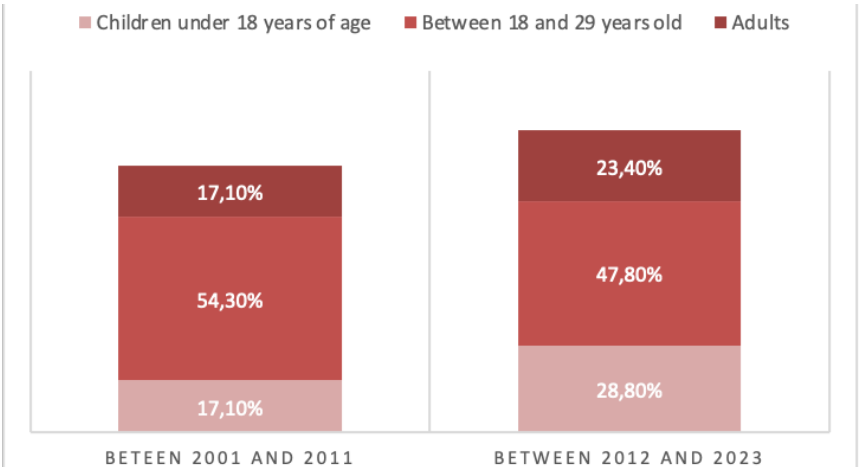
1 Complutense University of Madrid, Madrid, Spain, paulamnu@ucm.es

Similarly, in reference to what Neumann (2009) 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 Reinares, García and Vicente (2020) 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 (2016), 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 (2024) 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 percentage points compared to the period between 2001 and 2011.

Graph 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



Own elaboration based on Vicente (2024)

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 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 (2024).

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.

References

1. Bourekba, M.: ¿Por qué atrae el "Estado Islámico"? CIDOB (112), 1–5 (2015)
2. 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 accesed 2024/12/19
3. 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, 55–70 (2021)
4. Fanjul, M.L.: El mensaje persuasivo radical: yihadismo y redes sociales. Instituto Español de Estudios Estratégicos (115), 1–14 (2015)
5. González, M.I.: Perfil del terrorista y del reclutador yihadistas. Safe World, (2022) https://www.redsafeworld.net/post/perfiles-terroristas-reclutadores-yihadistas
6. Lejarza, E.: Terrorismo islamista en las redes – la yihad electrónica. Instituto Español de Estudios Estratégicos (100), 1–18 (2015)
7. López, S. et al.: Tendencias online de la propaganda yihadista. Análisis del caso español. Instituto Español de Estudios Estratégicos (85), 1–21 (2021)
8. Malatji, M., Tolah, A.: Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. AI Ethics, 1–28 (2024)
9. Montes, D.: A vueltas con el terrorismo e internet: hacia una definición de ciberterrorismo. Revista de Derecho UNED (27), 697–738 (2021)
10. 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)
11. Neumann, P.: Old & New Terrorism. In: Neumann, P. (ed.) Old & New Terrorism, pp. 14–48. Polity Press, Cambridge (2009)
12. Pisabarro, S.: La radicalización yihadista de los jóvenes en Europa a través de TikTok. Instituto Español de Estudios Estratégicos (2), 1–24 (2024)
13. Rapoport, D.C.: The Four Waves of Moderm Terrorism. In: Cronin, A., Ludes, J. (eds.) Attacking Terrorism, pp. 46–73 (2005)

14. Reinares, F., García, C., Vicente, Á.: Yihadismo y yihadistas en España. Quince años después del 11M. Real Instituto Elcano (159), 1–200 (2019)
15. 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/
16. Sánchez, L.M.: Repensando el concepto de ciberterrorismo. Instituto Español de Estudios Estratégicos (11), 1–15 (2021)
17. 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
18. Vicente, Á.: Radicalización yihadista en España: menores, espacios virtuales y la resonancia de conflictos internacionales. Real Instituto Elcano (31), 1–13 (2024)
19. Weimann, G. et al.: Generating Terror: The Risk of Generative AI Exploitation. Combating Terrorism Center At West Point (17), pp. 1–41 (2024)
20. Zelin, A.: The State of Global Jihad Online. A Qualitative, Quantitative, and Cross-Lingual Analysis. New America Foundation, 1–24 (2013)

ETHICAL CHALLENGES OF AI-DRIVEN TARGETED ADS FOR YOUTH

João Gonçalves[1] [0009-0008-4434-3909]

Keywords: AI, Ethics, Advertising, Youth.

Extended Abstract

The rise of Artificial Intelligence (AI) has brought multiple new technological advancements and business opportunities. One of these opportunities is AI-enhanced targeted advertising. Advertisements are everywhere—on TV, in stores, and on most websites. However, with more and younger users spending time on platforms that deploy these targeted ads multiple times, some ethical concerns arise.

AI can be hard to define, but to summarize "Artificial intelligence (AI) is a technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy." [1].

In advertisements, AI follows the same philosophy. It works by collecting user data, such as browsing history, search queries, geological locations, or time spent on certain content. Based on this information, the AI predicts what kind of ads a user is more likely to interact and engage with. This process is what makes AI enhanced targeting advertising so powerful and effective. As it continues to learn from users and their behaviors towards certain content, it refines ads to maximize engagement. [2]

AI-enhanced advertisements are highly effective, but with this effectiveness comes significant responsibility. One of the major concerns is data privacy, as AI to display these ads needs vast amounts of personal user data. These systems continuously collect user data often without users fully realizing its extent. [3].

Young users can be particularly vulnerable, as they do not fully realize and understand how these algorithms work and what data they are sharing. Without their knowledge, their digital footprint forms before they can truly understand data privacy. [4].

Moreover, companies often store this data long-term. Since different platforms use it, security concerns arise. Especially in cybersecurity where there could be a data leak exposing some of this user's personal information. Young users do not have the knowledge or awareness to be responsible for using tools to manage their online privacy. [5].

Additionally, algorithmic bias is a concern. Since AI models are only good as their training data. If the training data happens to contain biases, the AI will find these

1 Universidade Autónoma de Lisboa, Portugal, 30012425@students.ual.pt

patterns and replicate them. Sometimes even, amplify them. In advertising, this could mean the system targets certain groups of people towards a certain product, while a different group might be excluded completely from seeing that type of product. [6].

A good example is Facebook. A company placed an advertisement for science, technology, engineering, and math (STEM) career opportunity. The ad aimed to be gender-neutral, but due to Facebook's advertisement algorithm, this advertisement was shown more frequently to males than females. [7].

Unlike traditional advertising, where it is clear why companies place ads in a certain context, AI-enhanced advertising relies on complex algorithms that are difficult to understand. Users often do not understand why they are seeing a particular ad, making it harder to recognize patterns of influence. Every user has the right to understand why they are seeing a specific ad. This lack of transparency also makes it difficult to regulate or challenge AI-driven decisions, as companies using these technologies may not fully understand how their models prioritize content. [6] [8].

Most companies that are deploying these AI enhanced ad systems often do not fully understand how they operate. Since many businesses rely on third-party AI models or outsourced machine learning services. It is easy to understand why their knowledge is so limited. Often treating these AI's as "magic" that optimizes and boosts their engagement and sales. Deploying something and not bothering to fully understand, is ethically concerning as they do not understand the full capacities of the AI. This means that AI could display inappropriate content without clear accountability. Having a bigger impact on the younger audience. [8] [9].

Since AI cannot distinguish between ethical and unethical engagement, it will display a harmful or misleading ad if it performs well. The system will continue to promote this ad regardless of its consequences. This is of course an unintended effect on these companies' advertising strategies. With this, it becomes clear that we need strict safeguards to prevent showing this type of content. [10] [11].

Beyond this, we must also consider ad fraud. Since AI enhanced advertising systems are based on interactions. These systems are vulnerable to fraudulent activities such as bot traffic and click fraud. These fake interactions can manipulate and inflate engagement metrics gathered by the AI. Making advertisers believe their campaigns are performing well, while also distorting how the AI learns and optimizes future ads. As a result, AI may prioritize engagement from these automated systems since they generate more engagement. Making companies waste money on advertising for automated systems, which can affect the company's economy. [6].

To address these ethical concerns, companies need to take steps to make AI-enhanced targeted advertising safer and more transparent, for everyone, especially for young users.

Regarding data collection, there should be better data privacy rules. Platforms need to explain clearly, what data is used and collected. Data collection on young users should be opt-in, ensuring that no personal data is collected. Young users and their guardians should have total control over their data and its collection.

To reduce AI bias, developers need to check the training data used in the AI. It is important to make sure the data is unbiased and balanced. To mitigate further this issue, experts should inspect the training data and the AI itself every trimester to thoroughly analyze for potential biases.

Bot traffic and click fraud can be hard to spot, but by continuously analyzing traffic and their patterns, an AI system can spot irregularities and filter bot-generated clicks while also taking action to prevent showing those ads to automated systems.
As for transparency concerns, there should be a way to show users exactly what originated a certain decision, for example an ad for gaming mouse, could show that it was based on a recent search for a gaming mouse made by the user.

Finally, companies must take a sense of responsibility to continue to improve and make AI enhanced advertising safer and more ethical.

Conclusions

Advertising while also using a tool like AI is a great business opportunity, but this opportunity also comes with ethical challenges. As the use of AI in the advertising business increases, the need for responsible use grows. By actively using AI ethically, companies can build trust in their consumers and protect their reputation. Competitive markets, such as advertising, cannot ignore its advantages, as it can give an edge to the companies in a constantly changing market. Transparency and accountability are extremely important. Businesses and their users must understand how AI makes decisions, especially regarding young users, who might not grasp how their data is used. While using AI ethically, we should prevent dangerous or harmful content, protect user data, and reduce bias in ad targeting. Since AI systems can be tricked and influenced by bot traffic and click fraud, developers, and businesses need strong mechanisms to stop this fraud and ensure ads reach the desired audiences. By critically examining these ethical concerns and taking deliberate steps to address them, AI can create a better experience for everyone, not just as a business tool but as a way to improve advertising.

Acknowledgments: I would like to acknowledge the use of Grammarly for language improvements and proofreading.

References

1. Stryker, C., & Kavlakoglu, E. (2024, August 9). Artificial Intelligence. IBM. https://www.ibm.com/think/topics/artificial-intelligence
2. Kaput, M. (2024, January 22). AI in Advertising: Everything You Need to Know. Marketing AI Institute. https://www.marketingaiinstitute.com/blog/ai-in-advertising
3. Granados, A. (2024, November 15). AI and Personal Data: Balancing Convenience and Privacy Risks. Velaro. https://velaro.com/blog/the-privacy-paradox-of-ai-emerging-challenges-on-personal-data
4. Soni, Nitin. (2024). Digital footprints of the young: privacy concerns and awareness among teen social media users. https://www.researchgate.net/publication/387476950_Digital_footprints_of_the_young_privacy_concerns_and_awareness_among_teen_social_media_users
5. Del Valle, G. (2024, September 19). The FTC says social media companies can't be trusted to regulate themselves. The Verge. https://www.theverge.com/2024/9/19/24249073/ftc-data-retention-privacy-report-facebook-meta-youtube-reddit
6. Singh, P., Shukla, I., Singh, K., Dhingra, V., Dubey, S., Shakya, M., & Chawla, T. (2021). The role of AI in advertising effectiveness: A conceptual framework. https://www.neuroquantology.com/media/article_pdfs/08_Reserch_paper-1_kps.pdf

7. Lambrecht, A., & Tucker, C. (2019). Algorithmic Bias? An Empirical study of Apparent Gender-Based Discrimination in the display of STEM career ads. Management Science, 65(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093
8. Lu, S. (2020). Algorithmic opacity, private accountability, and corporate social disclosure in the age of artificial intelligence. Scholarship@Vanderbilt Law. https://scholarship.law.vanderbilt.edu/jetlaw/vol23/iss1/3/
9. Smith, B. (2024, July 30). Protecting the public from abusive AI-generated content. Microsoft on the Issues. https://blogs.microsoft.com/on-the-issues/2024/07/30/protecting-the-public-from-abusive-ai-generated-content/
10. Albrecht, J., Kitanidis, E., & Fetterman, A. J. (2022). Despite "super-human" performance, current LLMs are unsuited for decisions about ethics and safety. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2212.06295
11. Shin, D., & Jitkajornwanich, K. (2024). How Algorithms Promote Self-Radicalization: Audit of TikTok's Algorithm Using a Reverse Engineering Method. Social Science Computer Review, 42(4), 1020-1040. https://doi.org/10.1177/08944393231225547
12. Akter, S., Dwivedi, Y. K., Sajib, S., Biswas, K., Bandara, R. J., & Michael, K. (2022). Algorithmic bias in machine learning-based marketing models. Journal of Business Research, 144, 201–216. https://doi.org/10.1016/j.jbusres.2022.01.083
13. Grammarly. Grammarly Handbook. [Online] 2024. https://www.grammarly.com/

USING AI ASSISTANTS FOR ACADEMIC WRITING EXPLORING STUDENT PERSPECTIVES AND ETHICAL ASPECTS

Joana M. Matos[1] [0000-0002-1075-9075]
Adrian-Horia Dediu[2] [0000-0001-9119-7109]

Keywords: Academic Integrity, Ethical AI use, Preliminary Generation Z perspectives on AI.

Abstract

This paper explores current ethical challenges when using Artificial Intelligence (AI)-based tools for students and professors in the academic writing stages. 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.

Extended Abstract

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.

1Universidade Autónoma de Lisboa, Palácio Dos Condes Do Redondo, R. de Santa Marta 56,1169-023 Lisboa, Portugal; Nova School of Business and Economics, Rua da Holanda, n.º 1, 2775-405 Carcavelos, Portugal, jmatos@autonoma.pt
2 Universidade Autónoma de Lisboa, Palácio Dos Condes Do Redondo, R. de Santa Marta 56, 1169-023 Lisboa, Portugal, adediu@autonoma.pt

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.

Methodology

A survey with 10 questions was implemented in Moodle and administered to a group of 24 bachelor students. We categorize the questions into three thematic categories:

1. Questions with quantitative answers about the preference for various working methods used for academic work preparation,
2. Questions with qualitative answers exploring students' experience and perspectives on ethical AI use,
3. The extent to which students rely on AI in different stages of academic writing.

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.

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 (7). The repository contains the questions, the answers and the graphs with the average values for the numerical answers.

Concluding Remarks

Using AI opens new opportunities for enhancing academic writing productivity. However, we must consider ethical aspects before adopting AI. The results of our questionnaire show that students perceive AI as applicable but with well-defined limits. It is widely acceptable for research and programming but could affect academic integrity if inappropriately used in the final writing process. This article reinforces the need to improve the current standards and recommendations about the ethical use of AI in academic writing. Future research should explore institutional approaches to standardizing AI ethics in academic publishing. For the final version of this paper, we will present in detail the results and conclusions of our questionnaire. In addition, we intend to give a more detailed literature review, comparing our findings with similar works. We also intend to give several examples of using AI for quiz question design. In a future paper, we will share our experience with the methodology for producing raw Markdown-based slides with the help of AI for theoretical support and laboratory work in Programming courses. Acknowledgements. This text started with an intensive series of conversations with ChatGPT (8) models GPT 4o, o1, and o3-mini, 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 (9).

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

References

1. Academic integrity in the age of Artificial Intelligence (AI) authoring apps. Yeo, Marie. 03 2023, TESOL, Vol. 14.
2. IEEE Code of Ethics. [Online] https://www.ieee.org/about/corporate/governance/p7-8.html.
3. IEEE Code of Conduct. [Online] 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] https://www.acm.org/code-of-ethics.
5. Springer Nature, Editorial Policies: Artificial Intelligence (AI). [Online] https://www.springer.com/us/editorial-policies/artificial-intelligence--ai-/25428500.
6. Generation Z. Merriam-Webster.com Dictionary. [Online] Merriam-Webster. https://www.merriam-webster.com/dictionary/Generation%20Z.
7. Dediu, A.-H. GitHub. [Online] 2024. https://github.com/ahdediu/AI-Writing-Survey.git.
8. OpenAI. ChatGPT. [Online] 2023. https://chat.openai.com/.
9. Grammarly. Grammarly Handbook . [Online] 2024. https://www.grammarly.com/.

THE IMPACT OF MISOGYNISTIC, SEXIST AND ANTI-FEMINISTIC SPEECH IN SOCIAL MEDIA ON YOUTH AND THE ROLE OF AI.

Anna Maria Piskopani[1] [0000-1111-2222-3333], Liz Dowthwaite[2] [1111-2222-3333-4444], Richard Hyde[3] [2222-3333-4444-5555], Aislinn Bergin[4] [3333-4444-5555-6666], Boriana Koleva[5] [4444-5555-6666-7777], Hannah Heilbuth[6] [5555-6666-7777-8888], Nicholas Gervassis[7] [6666-7777-8888-9999], Dominic Price[8] [7777-8888-9999-10101010]

Keywords: Artificial Intelligence, Moderation, Misogyny.

Extended Abstract

Web 2.0 has been previously understood as an online public sphere that enables people to construct identities and communities, to influence official politics, and to promote democracy (Dahlgren, 2005). Thus, one of the new proposed digital rights is the right to participate in the digital online sphere. The European Declaration of Digital Rights advocates for the protection of the right to participate in the digital online sphere. Everyone should have access to a trustworthy, diverse, and multilingual digital environment. The justification is that access to diverse content contributes to a pluralistic public debate. The common element between the importance of the right to participate in the digital public sphere and the notion of well-being is the metaphor of "human flourishing" (Logan et al., 2023).

Due to gender patterns, there are gaps between how young men and women interact in the online spaces. Young men are more likely to share content online, to look for information, read news about politics, while young women more often go online for social interaction and relationship maintenance (Bode, 2017). Additionally, young women are more likely to feel that when they share their views, they are exposed to bullying, harassment and hate speech online. 60% of women (compared to 31% of men) feel that their gender is the reason for the reactions to their views (UNDP, 2021).

The rise of the misogynistic, sexist and anti-feministic online speech in social media in the recent years poses the question of whether youths (young people aged from 18 to 24) are affected by this, how so, and the role that AI algorithms and new AI moderation techniques play on this. Researchers recently conducted a survey in the

1 University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom
2 University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom
3 University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom
4 University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom
5 University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom
6 University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom
7 University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom
8 University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom

UK to examine the impact of anti-gendered speech. The survey revealed that women primarily fear of being trolled or targeted by misogynists when they express themselves online and share photos (Stevens et al., 2024). There is evidence that algorithms in social media privilege extreme material over less provocative one, such as misogynistic speech viewed by young people, provoking a similar impact on their offline behaviour to girls (Cowood, 2024). Social media platforms use algorithms to create users' profiles and keep them engaged in content (e.g. videos on YouTube), as advertisers pay based on viewers' attention (Zuboff, 2019, Amnesty International, 2023). Young users can view somewhat neutral content and "can suddenly find themselves immersed in more extreme or harmful material" (Spring, 2024, Hopkins, 2024) as the algorithm identifies that content as their interests will actively amplify and direct that content (Regehr et al, 2024). Additionally, AI generated tools have added new ways of women's online abuse and harassment like sexualised deepfakes (Dunn, 2021). There is a legitimate fear that anti-feministic, sexist and misogynistic content normalises certain ways of thinking and increases inequalities between groups (Theilen, 2023) and is an attempt to silence women and increase gender inequality with severe political and social implications.

In this paper, we are going to present part of our findings of our research project "Gendered exclusion and wellbeing on the internet" (University of Nottingham, Horizon Institute of Digital Economy, Welfare Campaign) which focuses on the internet as a technology which affects people's wellbeing. Our project key aims are to explore how the misogynistic, sexist and anti-feministic speech creates barriers that prevent people from sharing opinions and thoughts in social media and from generally feeling able to participate on online platforms, and how this directly affects their wellbeing. We conduct quantitative research on social media users. Additionally, as the University of Nottingham attracts students from multiple cultural and ethnic backgrounds, we conduct qualitative research on University of Nottingham students and staff. We are currently collecting responses, and our overall aim will be for online survey 100-150 and interviews 50.

Our first aim was to investigate: A) how they are impacted by the use and rise of misogynistic, antifeminist, hostile and manosphere online speech, B) whether they have experienced gendered digital exclusion or felt that their gender made it more difficult for them to participate due to their fear of being attacked. Additionally, we focused on the role that AI technologies play in this context. Given that Generative AI tools can help online users to create misogynistic, sexist and anti-feministic content, we investigated whether young people have encountered this AI content (such as sexualised deepfakes videos and images, AI generated images that enforce gender stereotypes and AI chatbots making misogynistic jokes or comments) and whether they believe that they can recognise AI generated content.

As AI moderation services on social media can automatically make moderation decisions -- refusing, approving or escalating content – and they continuously learn from their choices, we asked online users, University of Nottingham students and staff, whether algorithms on certain platforms show them content that they regard as misogynistic, sexist or anti-feminist. We also asked them if they ever felt that content, they have posted was detected by AI algorithms as misogynistic, sexist or antifeminist even if it was not or if it should be detected and it was not. Finally, we shared these expressed views and

concerns with various professional stakeholders, such as social media and tech industry developers, civic organisations, policymakers about effective ways to mitigate the phenomenon. The aim of this study is not only to capture the impact of such speech on mental health and online participation but also to raise public awareness about the extent of the phenomenon and its social harms, suggest ways to mitigate it, and provide resources to educators and individuals to address it and discuss openly about these issues.
In this paper we will present our research findings focused on the young people aged from 18 to 24 in comparison with the other population and search whether young people are less or more likely to take advantage of the opportunities raised by communication in social media and confront the new AI challenges.

Acknowledgments: The work was supported by the Engineering and Physical Sciences Research Council "Horizon Trusted Data Driven Products" Award (grant number EP/T022493/1).

References

1. Amnesty International Homepage, "I feel exposed". Caught in Tik Tok surveillance web, https://www.amnesty.org/en/documents/pol40/7349/2023/en/, last accessed 2025/3/20
2. Bode, L.: Closing the gap: gender parity in political engagement on social media, Information, Communication & Society 20 (4), 587-603 (2017)
1. Cowood, F. (2024, June 20). The rise of the aggro-rithm: Can misogynistic content be stopped? The Guardian. https://www.theguardian.com/a-matter-of-connection/article/2024/jun/20/the-rise-of-the-algorithms-can-misogynistic-content-be-stopped-ai, last accessed 2025/02/01.
2. Dahlgren, P.: The Internet, Public Spheres, and Political Communication: Dispersion and Deliberation. Political Communication, 22(2), 147–162 (2005)
3. Dunn, S. Women Not politicians, are targeted most often by deepfake videos, Centre for International Governance Innovation, https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/, last accessed 2025/02/01
4. Hopkins, S. (2024, September 2). The Link Between Social Media Algorithms and Online Radicalisation. Byline Times. https://bylinetimes.com/2024/09/02/social-media-algorithms-and-online-radicalisation/, last accessed 2025/02/01
5. Logan, A. C., Berman, B. M., & Prescott, S. L. Vitality Revisited: The Evolving Concept of Flourishing and Its Relevance to Personal and Public Health. International Journal of Environmental Research and Public Health, 20(6), 5065-5080 (2023)
6. Regehr, K., Shaughnessy, C., Zhao, M., Shaughnessy. N., Safer scrolling: How algorithms popularise and gamify online hate and misogyny for young people. UCL IOE, University of Kent, Align Platform. https://www.alignplatform.org/resources/safter-scrolling-how-algorithms-popularise-and-gamify-online-hate-and-misogyny-young, last accessed 2025/02/01
7. Spring, M. (2024, September 2). Social media: Why algorithms show violence to boys. BBC News. https://www.bbc.com/news/articles/c4gdqzxypdzo, last accessed 2025/02/01
8. Stevens, F., Enock, F., Sippy, T., Bright, J., Cross, M., Johansson, P., Wajcman, J., Margetts H., (2024, March). Understanding gender differences in experiences and concerns surrounding online harms: A nationally representative survey of UK adults. The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2024-03/understanding_gender_differences_in_experiences_and_concerns_surrounding_online_harms_-_a_nationally_representative_survey_of_uk_adults.pdf, last accessed 2025/02/01
9. Theilen J., Article 20: Digital inequalities and the promise of equality before the law. In Giannopoulou A., Digital Rights are Charter Rights – Essay Series, Digital Freedom Fund, https://digitalfreedomfund.org/digital-rights-are-charter-rights-essay-series/, last accessed 2025/02/01

10. UNDP (2021), Civic participation of youth in a digital world – Europe and Central Asia, htpp://www.undp.org/eurasia/publications/civic-participation-youth-digital-world, last accessed 2025/02/01
11. Zuboff. S. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Profile Books. London (2019)

ETHICAL CONSIDERATIONS AND GOVERNANCE FRAMEWORKS IN THE DIGITAL AGE

Maria Albertina Rodrigues[1] [2A16-8A66-42BB], Ana Cristina Saraiva[2] [9E1F-3F96-35A3]

Keywords: AI Ethics, Youth and Social Media, AI Governance.

Introduction

The integration of artificial intelligence (AI) in social media platforms predominantly used by young people has introduced some ethical challenges that require urgent attention. AI-driven algorithms influence content selection, shape digital interactions, and impact youth behavior in ways that raise concerns about privacy, mental health, algorithmic biases, and digital literacy. These concerns align with global initiatives such as the Global Digital Compact, which advocates for a human-centered, rights-based, and inclusive digital future.

This paper aims to explore the ethical considerations surrounding AI deployment in youth-oriented social media environments, emphasizing governance structures and regulatory frameworks that ensure AI-driven technologies prioritize youth well-being. As organizations like UNA Portugal advocate for responsible digital policies, this discussion becomes particularly relevant within both the Portuguese and global digital landscapes. Additionally, this paper draws on research concerning the transformation of ICT values during uncertain times (Saraiva et al., 2021), analyzing how crises—such as the COVID-19 pandemic—accelerate digital transformation and reshape ethical norms.

The three main research questions of this proposal are:
(1) What are the main ethical implications of AI in Social Media for young users?, (2) Which is the role of governance in the AI ethics?, (3) Which is the lessons learned from a case study of digital transformation?.

These questions will guide the paper´s exploration of AI ethics, policy frameworks, and the evolving role of AI in youth-oriented digital environments.

The Ethical Implications of AI in Social Media for Young Users

AI systems are the backbone of social media platforms, influencing user experiences through content recommendation algorithms, targeted advertising, and moderation tools. While these technologies optimize engagement and personalization, they also

1 Universidade Europeia; CETRAD; ISCAL/ Polytechnic University of Lisbon, Portugal; APEE, Lisboa, Portugal
2 APEE, Lisboa, Portugal, anasaraiva@apee.com

present ethical risks, particularly for young users who may be more susceptible to digital harm. The most critical concerns include:

AI-Driven Content moderation and Its Impact on Youth Well-Being

Recommendation algorithms optimize content for engagement, often amplifying emotionally charged material. Research suggests that prolonged exposure to extreme content can negatively affect youth mental health, contributing to anxiety, depression, and harmful self-perception (Tavares, 2003). This paper examines how ethical AI design can mitigate these risks and promote youth digital well-being.

Privacy and Data Protection in Social Networks

Young users often share personal data without a full understanding of AI tracking and profiling mechanisms. AI systems collect vast amounts of personal information to refine targeting strategies, raising concerns about user autonomy and data governance. The Global Digital Compact calls for stronger policies that safeguard digital rights, emphasizing the role of regulations such as the EU's General Data Protection Regulation (GDPR) and Portugal's digital policies in ensuring responsible AI utilization.

Algorithmic Bias, Discrimination, and Representation in AI

AI models inherit biases from training data, leading to discriminatory outcomes in content recommendations and moderation. Marginalized youth communities are disproportionately affected, limiting their visibility and access to digital opportunities. This research investigates how ethical AI frameworks can address algorithmic discrimination and ensure fair representation in online spaces (Saraiva et al., 2021).

Ethical Considerations in AI-Enhanced Targeted Advertising for Young Audiences

AI-driven advertising systems exploit behavioral data to personalize ads, sometimes blurring the line between marketing and coercion. The COVID-19 pandemic saw a rise in AI-powered e-commerce and digital advertising, raising questions about ethical consumption and youth autonomy (Saraiva, 2022). This paper assesses the impact of AI-enhanced advertising on youth decision-making and explores potential policy interventions.

Transparency and Explainability of AI Systems in Youth-Oriented Platforms

Many AI-driven systems operate as "black boxes", where users don´t have much clarity on how content is selected, moderated, or restricted. Youth may struggle to understand why certain posts appear in their feeds or why they are subjected to algorithmic decisions. Transparency and explainability in AI governance are crucial to fostering trust and digital literacy. This research explores methods for designing AI systems that prioritize fairness, interpretability, and accountability.

The Role of Governance in AI Ethics

The governance of AI-driven social networks requires a multi-stakeholder approach involving governments, policymakers, tech companies, educators, and civil soci-

ety organizations. The Global Digital Compact promotes human-centered AI policies, while national initiatives, such as UNA Portugal, contribute to discussions on AI ethics.

Regulatory Considerations

While the GDPR provides a robust legal framework for digital rights protection in Europe, additional measures are required to address emerging AI-related risks. This study reviews existing AI governance frameworks and identifies regulatory gaps that need to be addressed to enhance AI accountability and compliance.

Educational Initiatives for Digital Literacy

Ensuring ethical AI development requires robust digital literacy programs that empower young users to critically engage with social media environments. Schools, universities, and NGOs play a fundamental role in equipping youth with the knowledge and skills needed to navigate the digital landscape responsibly. This paper examines which are the best practices in digital literacy education and their impact on AI awareness among young users.

AI and Civic Engagement Among Young People

AI-driven platforms have the potential to foster youth civic engagement by providing access to diverse information sources and enabling digital activism. However, concerns about algorithmic bias, political polarization, and misinformation challenge AI's role in democracy. This research tries to investigate how AI can be leveraged to strengthen civic participation while minimizing misinformation risks.

Lessons Learned: A Digital Transformation Case Study

The COVID-19 pandemic accelerated AI adoption across multiple domains, reshaping ethical considerations and digital governance priorities (Saraiva et al., 2021). Key lessons include:

Adjustments in ethical values related to AI consumption: The pandemic highlighted concerns about digital security, privacy, and mental health, leading to greater demand for transparency in AI systems.

AI in distance learning and education: Online education became a necessity, with e-learning platforms playing a central role. However, concerns about accessibility, surveillance, and algorithmic biases emerged.

The need for stronger AI governance: Increased reliance on digital platforms during the pandemic highlighted the need for ethical AI policies that protect youth users from exploitation and misinformation.

Methodology

To ensure a robust empirical foundation, this paper employs a mixed-method approach, combining quantitative data analysis and qualitative research:

Quantitative Analysis

Survey on AI Awareness and Digital Literacy: A large-scale survey of Portuguese youth (ages 13-24) will assess their understanding of AI, privacy concerns, and social media experiences. This survey will include:
Sentiment Analysis of AI-Generated Content: Using QuestionPro Software, this analysis tries to understand the sentiment trends in social media posts directed at youth.
Algorithmic Bias Testing: Using AI models trained with different datasets, this study will evaluate disparities in content recommendations based on gender, ethnicity, and socio-economic status.

Qualitative Analysis

Expert Interviews: AI ethicists, policymakers, and educators will provide insights into governance challenges and AI literacy initiatives. This analysis will include:
Policy Review: Comparative analysis of AI governance frameworks, with a focus on the Global Digital Compact, GDPR, and Portuguese digital policies.

Concluding Remarks

As AI continues to shape youth minds on social media, ethical considerations must remain at the forefront of technological development and governance. This research fosters interdisciplinary discussions that bridge AI ethics, policy, education, and youth advocacy. By integrating insights from global initiatives like the Global Digital Compact and local efforts by UNA Portugal, this paper aims to contribute to a more inclusive and responsible digital future.

References

1. European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act) and amending certain Union legislative acts. COM(2021) 206 final
2. Saraiva, A. & Silva, N. (2021). COVID-19 impact on online education – Opportunities and Challenges in a SWOT Analysis. International Journal of Research in E-Learning. 7(2): 1-18
3. Saraiva, A., Tavares, L. V. & Silva, N. (2022). Emerging values in ICT during uncertain times: the case of COVID- 19. ETHICOMP 2022 Conference. 26 to 28 July, Turku, Finland
4. Tavares, L., Ferreira, J., Saraiva, A. & Sá, A. (2023). A Multicriteria Evaluation of Local E-Government in Terms of Citizen Satisfaction. International Business Research. 16(12): 20-40
5. United Nations (2023). Global Digital Compact – Background and Draft Elements

SOCIAL MEDIA AS LEARNING ECOSYSTEMS: ETHICAL CONSIDERATIONS AND THE ROLE OF ARTIFICIAL INTELLIGENCE IN YOUTH SKILL DEVELOPMENT

Luís Valadares Tavares[1], Nuno Silva[2], Joana Cruz[3], Sara Cruz[4], Paula Rodrigues[5], Alexandre Ricardo[6], Mohsen Motamedi[7], Beatriz Santos[8]

Keywords: AI-driven education, decision-making, social media, youth learning ecosystems.

Extended Abstract

Social media platforms have revolutionized how young adults engage with information, transitioning from traditional learning environments to digital and interactive ecosystems. While social media has often been associated with informal communication and entertainment, its role in education and skill development is expanding. Integrating Artificial Intelligence (AI) in social media learning ecosystems provides an opportunity to structure and personalize learning experiences, bridging the gap between passive content consumption and active knowledge acquisition (Holmes et al., 2019). This paper examines how youth engage with social media for skill development, AI's role in shaping these learning pathways, and the ethical considerations associated with algorithm-driven learning environments. AI-driven education is transforming traditional pedagogical models by offering personalized and adaptive learning experiences. Ouyang and Jiao (2021) categorize AI applications in education into three paradigms: AI-directed, AI-supported, and AI-empowered learning. AI-directed learning relies on automated systems to guide educational content, AI-supported learning enhances traditional learning methods with intelligent systems, and AI-empowered learning allows learners to take an active role in shaping their educational pathways. These models demonstrate how AI can facilitate the transition from unstructured to structured learning on social media platforms (Ouyang & Jiao, 2021).

AI and Personalized Learning Pathways

The ability of AI to personalize learning experiences is one of its most promising aspects. AI-driven recommendation algorithms analyze user interactions to tailor edu

1 COMEGI, Lusíada University, Portugal, lvt@lis.ulusiada.pt
2 COMEGI, Lusíada University, Portugal
3 COMEGI, Lusíada University, Portugal
4 COMEGI, Lusíada University, Portugal
5 COMEGI, Lusíada University, Portugal
6 COMEGI, Lusíada University, Portugal
7 COMEGI, Lusíada University, Portugal
8 COMEGI, Lusíada University, Portugal

cational content, making learning more relevant and engaging (Zhao, 2022). Studies on non-formal education highlight the benefits of personalized learning, particularly in digital environments where learners can set their own pace and goals (Widodo & Nusantara, 2020). However, while AI-driven personalization offers significant advantages, it also raises concerns about the reinforcement of pre-existing biases. Algorithmic bias may lead to a narrowing of content exposure, restricting learners to perspectives aligned with their previous interactions (Holmes et al., 2019). Ethical considerations are central to AI-enhanced education. The increasing reliance on algorithm-driven learning pathways raises concerns about data privacy, transparency, and content curation biases (Souto-Otero, 2021). AI models often rely on historical data to make predictions, which can perpetuate existing inequalities in access to quality education. Addressing these challenges requires greater transparency in AI recommendation systems, ensuring learners have control over their learning experiences and are exposed to diverse perspectives (Bouton et al., 2021).

Knowledge Sharing and Peer Learning

Social network technologies (SNTs) play a crucial role in facilitating knowledge sharing among students. Studies indicate that students use SNTs primarily for exchanging study materials rather than for collaborative knowledge construction (Bouton et al., 2021). Knowledge-sharing practices on social media platforms are often informal, yet they significantly contribute to students' learning experiences. Peer-to-peer learning through social media enables students to access a broader range of perspectives and educational resources (Asterhan & Bouton, 2017). However, the unstructured nature of knowledge sharing on social media presents challenges in ensuring information accuracy and reliability. Unlike formal education environments where instructors curate learning materials, social media learning relies on user-generated content, which can vary in credibility. AI has the potential to mitigate these challenges by verifying and prioritizing high-quality educational resources (Holmes et al., 2019). Implementing AI-based credibility assessments can help ensure that learners receive accurate and reliable information while maintaining the benefits of peer collaboration (Zhao, 2022).

The Role of AI in Non-Formal Education

Non-formal education (NFE) complements traditional learning by providing flexible and accessible educational opportunities. AI-driven interventions can enhance NFE by offering adaptive learning experiences tailored to individual learners' needs (Widodo & Nusantara, 2020). Research highlights that students engaged in NFE programs often benefit from AI-powered tutoring systems that provide personalized feedback and learning recommendations (Souto-Otero, 2021). The validation of NFE within formal education systems remains a challenge. Many institutions lack mechanisms to recognize skills acquired through informal learning environments (Souto-Otero, 2021). AI-driven assessment tools can bridge this gap by evaluating learners' competencies and mapping them to standardized educational frameworks. By integrating AI into credentialing

processes, institutions can provide greater recognition for knowledge and skills acquired through social media and non-traditional learning pathways (Widodo & Nusantara, 2020).

Challenges and Ethical Considerations

Despite the numerous benefits AI brings to education, several ethical concerns must be addressed. Algorithmic transparency remains a major issue, as learners often have limited insight into how AI systems determine their learning pathways (Holmes et al., 2019). The use of AI can also contribute to the spread of misinformation, as these models carry the risk of generating inaccurate content, which often exacerbates discrimination and stereotypes, especially against minority groups (Shalevska, 2024). Ethical AI frameworks should prioritize user autonomy, ensuring that students have control over their learning experiences and can critically engage with AI-driven recommendations. Additionally, the risk of AI reinforcing socio-economic disparities in education cannot be overlooked. Research indicates that students from disadvantaged backgrounds may have limited access to AI-enhanced learning tools, exacerbating existing educational inequalities (Ouyang & Jiao, 2021). Policymakers and educators must work collaboratively to develop equitable AI-driven education policies that ensure inclusive access to digital learning resources (Souto-Otero, 2021).

Conclusion

The integration of AI into social media learning ecosystems presents both opportunities and challenges for skill development among young adults. AI-driven personalization can enhance learning experiences, but concerns regarding algorithmic bias, ethical transparency, and equitable access remain critical considerations. By fostering interdisciplinary collaboration between educators, policymakers, and technology developers, AI-enhanced education can be designed to prioritize fairness, inclusivity, and learner autonomy. Future research should explore longitudinal studies on AI-driven learning interventions, assessing their long-term impact on educational equity and knowledge acquisition.

References

1. Asterhan, C.S.C. & Bouton, E. (2017). Teenage peer-to-peer knowledge sharing through social networks. Computers & Education, 110, 16-34.
2. Bouton, E., Tal, S.B., & Asterhan, C.S.C. (2021). Students, social network technology, and learning in higher education: Visions of collaborative knowledge construction vs. the reality of knowledge sharing. The Internet and Higher Education, 49, 100787.
3. Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign.
4. Ouyang, F. & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. Computers and Education Artificial Intelligence, 2, 100020.
5. Shalevska, E. (2024) Human Rights in the Age of AI: Understanding the Risks, Ethical Dilemmas, and the Role of Education in Mitigating Threats. Journal of Legal and Political Education, 1(2): 38-52. https://doi.org/10.47305/JLPE2412038sh.

6. Souto-Otero, M. (2021). Validation of non-formal and informal learning in formal educa. tion: Covert and overt. European Journal of Education, 56(3). https://doi.org/10.1111/ ejed.12464
7. Widodo, & Nusantara, W. (2020). Analysis of non-formal education (NFE) needs in schools. Journal of Nonformal Education, 6(1), 69-76.
8. Zhao, J. (2022). Influence of knowledge sharing on students' learning ability under the background of 5G+AI. International Journal of Emerging Technologies in Learning, 17(1), 133-145.

IMPACT OF PERSONAL GROWTH ON ADOLESCENTS' DECISIONS TO STUDY ON SOCIAL MEDIA

Nuno Silva[1] [0000-0003-0157-0710], Isabel Alvarez[2] [0000-0003-4381-6328],
Luis Valadares Tavares[3] [0000-0001-5727-1264], Elisa Cunha[4] [0000-0001-6415-7160]

Keywords: Adolescents, AI-driven study, decision-making, social media

Extended Abstract

Nowadays social media offers an important source of learning. In particular, social networks provide learning communities and discussions while adding Artificial Intelligence (AI) tools to facilitate learning and knowledge sharing. This paper takes a theoretical approach to adolescence in terms of its choices and what then leads to practice after the age of 17. So far, the decision-making styles and moral development of adolescents in the context of social networks involve complex interactions between cognitive, emotional, and social factors [1][2].

Adolescence is an important transition in life. It is a time of rapid development, spending more time with friends and less with family, trying out different personal styles, different hobbies, all signs of growth towards adulthood [3]. Adolescents are figuring out who they are and what they believe, besides adjusting to a changing body.

Social media is one of the most enthusiastic activities online where adolescents are considered the heaviest users [4]. They concluded that there should be guidelines in choosing social media contents used as classroom teaching materials and as a tool to enhance learning.

The way adolescents make decisions is one of the behavioral indicators of autonomy [5] with autonomy being defined as the capacity for self-direction and self-regulation of individuals [6]. As per [7] the influence of parental control on adolescent decision making has two dimensions: behavioral (refers to attempts to control adolescent behavior) and psychological (refers to parental control that interferes with adolescent emotional and psychological development using manipulation, criticism, among others).

The results of [8] research evidence that both parental control practices and adolescent temperamental dimensions are associated with decision-making, concluding that in middle and late adolescence, it is assumed beneficial for adolescent welfare to increase autonomy in decision-making on issues of personal and multifaceted domain.

1 COMEGI, Lusíada University, Portugal, nsas@ulusiada.pt
2 COMEGI, Lusíada University, Portugal
3 COMEGI, Lusíada University, Portugal
4 COMEGI, Lusíada University, Portugal

[9] investigated the relationships between the network characteristics of Year 10 students in Australia with their decision to enroll in a science course in the last year of secondary school, supporting the validity of network theory in the context of science interest, as central indicators apparently play an influential role in the network.

Frequently adolescents are often portrayed as passive receivers of social influence, where peers easily sway their decisions. However, following the results of a study [10], adolescents are also motivated to gain information about the preferences of specific peers, and thus play a role in selecting the sources that may inform their decisions, demonstrating a preference for their friends over non-friends, as well as for peers who were perceived as trustworthy.

The study of the development of the teenager's personality has been one of the most studied topics in Psychology identifying the most relevant factors to build up their key five trait trajectories: openness, conscientiousness, extraversion, agreeableness, neuroticism [11].

However, most of the studies have emphasized the importance of the main sources of influence to develop their self-estimation and self-development such as family, social group and social networks [12] but rather in terms of helping to improve their own ability to make decisions for their personal, professional and social achievement. In this research plan, models of Decision Sciences are used to support strategies to help teenagers to improve such ability considering multiple attributes, uncertainty and their lifelong future [13] following previous results recently published [14].

The authors believe that this approach based on the reinforcement of rational ability of teenagers to make their own decisions for their own good can be quite effective if compared to mimetic and persuasive approaches although much less studied.

Moral decision making is a complex process involving many components. Adolescents begin to make choices within a behavioral ecology where for many of them there are emerging issues in the decision-making process [15], and most of them operate at the Kohlberg's conventional level, where societal norms and relationships play a central role in moral reasoning. Cognitive growth, such as the development of formal operational thinking (as per Piaget), supports this progression by enabling adolescents to consider multiple perspectives.

Morally developed adolescents are more likely to use social networks responsibly for studying, with the ability to critically evaluate information found on social networks, to engage in respectful and constructive interactions, creating a positive environment for collaborative learning helping them to discern between reliable and unreliable sources for their studies.

Concerning the development of independence and identity, there is a need for a deeper understanding of the impact of social media on adolescent moral development and the crucial role of parents in guiding adolescents in responsible social media usage [16].

The rise of AI's effects on moral development of adolescents and counseling practice for learning and enhance to higher levels of ethical reasoning is very limited. In this regard, AI, which could be described as a computer system that behaves intelligently; when it was first introduced, was only used in some specialized domains, but with advances in machine learning techniques their applications have broadened [17].

If AI systems are designed in a more transparent way and thus are more explainable from the user's point of view [18], these systems will be considered more reliable and contribute to the users' satisfaction. [19] referred that the level of system transparency affects not only trust but also the results obtained of the system.

So, we believe that the transparency and explainability of AI systems (XAI) in youth-oriented social networks are critical for building trust, ensuring ethical use, and fostering digital literacy among adolescents. Many AI-powered apps used in social networks function as "black boxes," making it difficult for users to understand how decisions (e.g. recommendations or content moderation) are made. Therefore, based on the work done [20] in the medical domain, we believe that it is necessary to integrate the study needs of adolescents and build user-centered explainable XAI design practices in the domain of social networks. With this new approach of XAI systems, adolescents could be compelled to use social networks to study as a further source to their academic preparation.

Finally, important research questions emerged from this research:

- How and what do adolescents learn on social networks?
- Is decision style related to personal growth?
- What about transparency and explainability of AI systems in adolescent-oriented social networks?

References

1 Elwadhi, S. (2024). The Impact of Social Media on the Decision Making of Youth: A Survey-Based Analysis. Innovative Research Thoughts, 10(2), 57–69, DOI: 10.36676/irt.v10.i2.08.

2 Jacob, P. & Agarwal, S. (2024). Impact of Social Media on the Decision-making Process of Student. **http://dx.doi.org/10.2139/ssrn.4927276**

3 Yang. Z., & Laroche, M. (2011). Parental responsiveness and Adolescent susceptibility to peer influence: A cross-cultural investigation. Journal of Business Research, 64(9), 979-987

4 Nuñez-Rola, C. & Ruta-Canayong,(2019). N. Social Media Influences to Teenagers, International Journal of Research Science & Management, DOI: 10.5281/zenodo.3260717

5 Zimmer-Gembeck, M. & Collins, W. (2003). Autonomy development during adolescence. In G. R. Adams, & M. D. Berzonsky (Eds.), Blackwell handbook of adolescence (pp. 175–204). Malden, MA: Blackwell Publishing

6 Hill, J., & Holmbeck, G. (1986). Attachment and autonomy during adolescence. Annals of Child Development, 3,145–189.

7 Barber, B. K. (1996). Parental psychological control: Revisiting a neglected construct. Child Development, 67(6), 3296–3319. **https://doi.org/10.2307/1131780**

8 Pérez, J. Carola & Cumsille, Patricio (2012). Adolescent temperament and Parental Control in the development of the adolescent decision making in a Chilean sample. Journal of Adolescence 35 659-669

9 Sachisthal, M.; Jansen, B.; Dalege, J. & Raijmakers, M.E.J. (2020). Relating Teenagers' Science interest work characteristics to later science course enrolment: An analysis of Australian PISA 2006 and longitudinal Surveys of Australian Youth Data. Australian Journal of Education Vol. 64(3) 264-281

10 Slagter, S.K.; Gradassi, A.; Duijvenvoorde, A.C.K.von; Wouter van den Bos (2023). Identifying Who Adolescents Prefer As Source of Information Within Their Social Network, Scientific Reports doi.org/10.1038/s41598-023-46994-0

11. Tetzner, Julia; Becker, Michael & Bihle, Lilly-Marlen (2023). "Personality development in adolescence: Examining big five trait trajectories in differential learning environments", European Journal of Personality, Vol. 37(6) 744–764, DOI: 10.1177/08902070221121178
12. Lajnef, Karina. (2023). The effect of social media influencers' on teenagers Behavior: an empirical study using cognitive map technique, Current Psychology (2023) 42:19364–19377, https://doi.org/10.1007/s12144-023-04273-1
13. Reyna, V.F. & Farley, F. (2006). Risk and Rationality in Adolescent Decision Making: Implications for Theory, Practice, and Public Policy. Psychological Science in the Public Interest, 7(1), 1-44. https://doi.org/10.1111/j.1529-1006.2006.00026.x
14. Pragathi, Subramanian; Narayanamoorthy, Samayan; Pamucar, Dragan & Kang, Daekook. (2025). An Adaptive Decision-Making System for Behavior Analysis Among Young Adults, Cognitive Computation (2025) 17:25 https://doi.org/10.1007/s12559-024-10372-3
15. Susanu, N. (2023). Moral Development in Adolescents. New Trends in Psychology, Vol. 5, No. 1/2023, pp. 30-34
16. Hermansyah, D., Syaharuddin, & Amir, L. S. (2024). The influence of social media usage on the ethical and moral behavior of high school adolescents. International Conference on Education, Teacher Training, and Professional Development, 35-44.
17. Ha, T., Sah, Y., Park, Y., Lee, S. (2022). Examining the effects of power Status of an Explainable Artificial Intelligence System on Users' Perceptions, 41(5), 946-958.
18. Ward, B., Bhati, D., Neha, F., & Guercio, A. (2024). Analyzing the Impact of AI Tools on Student Study Habits and Academic Performance. arXiv. https://arxiv.org/html/2412.02166v1
19. Sadler, G., H. Battiste, N. Ho, L. Hoffmann, W. Johnson, R. Shively, … D. Smith. 2016. "Effects of Transparency on Pilot Trust and Agreement in the Autonomous Constrained Flight Planner." Paper presented at the 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC).
20. He, X., Hong, Y., Zheng, X. & Zhang, Y. (2023). What Are the Users' Needs? Design of a User-Centered Explainable Artificial Intelligence Diagnostic Systems, International Journal of Human-Computer Interaction, 39(7), 1519-1542.

ETHICOMP 2025 is a multitrack conference that brings together diverse sessions at the intersection of ethics, technology, society, and education. Designed to foster interdisciplinary dialogue, the program offers both thematic depth and crossdomain perspectives. Domain experts chair each track, reflecting the breadth of ethical challenges in contemporary information and communication technology (ICT)—from algorithmic fairness and data protection to AI governance, digital pedagogy, and cyberethics. To reflect the diverse ethical questions raised by digital technologies, ETHICOMP 2025 is organized around the following topical areas:

- Let's Reevaluate Cybersecurity Best Practices
- Ethical Challenges of Human Relations in Education in the AI Era
- Gender and Future Digital Technologies
- The Normative Challenges of AI Literacy
- (De)stigmatizing Digital Interactions and Technologies
- Artificial Intelligence: Interpretability and Measurement of its Underlying Ethics
- Values: The Basis for Responsible Technology and Digital Interactions
- How is AI Changing Us? How Does it Change the World?
- Ethical Challenges for Business in a Digital World
- AI, Social Networks, and Their Influence on Youth

UAL
AUTÓNOMA
CULTURA
edições