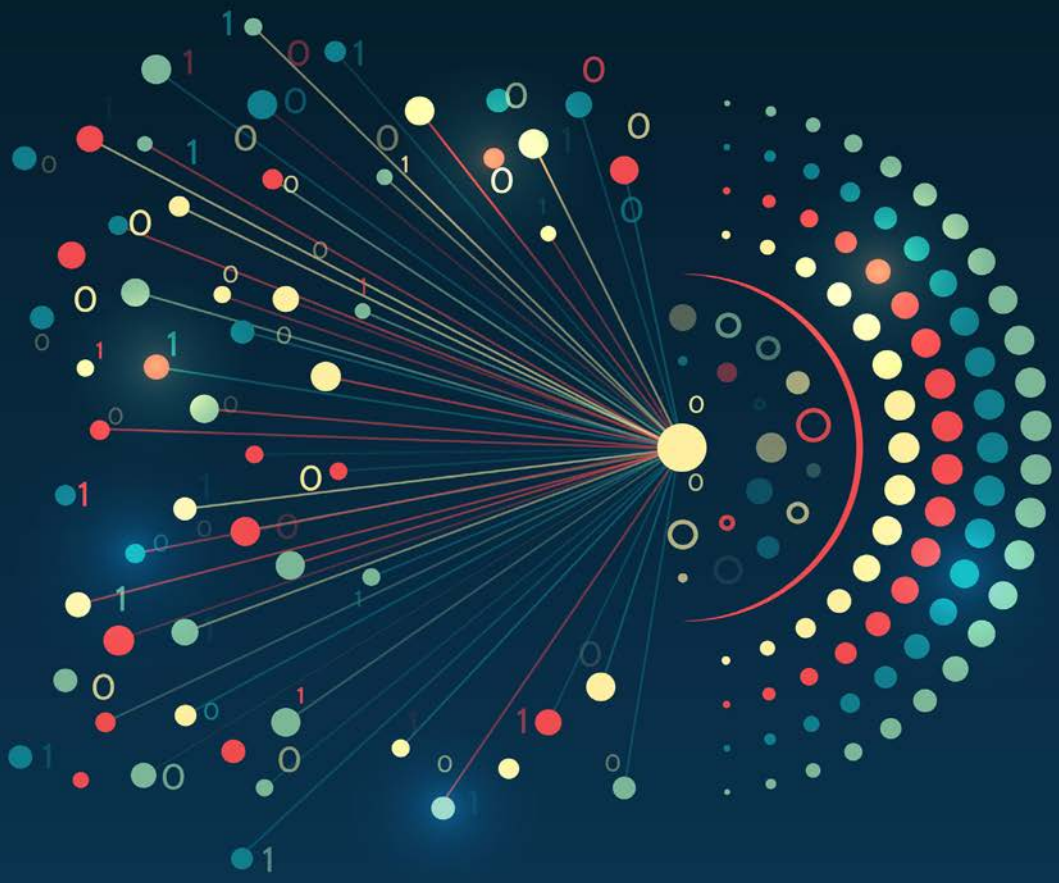
(Ref)using AI



21th INTERNATIONAL CONFERENCE
ON THE ETHICAL AND SOCIAL IMPACTS OF ICT

Smart Ethics in the Digital World
Proceedings of the ETHICOMP 2024

Edited by
MARIO ARIAS-OLIVA
JORGE PELEGRÍN-BORONDO
KIYOSHI MURATA
ANA MARÍA LARA PALMA
MANUEL OLLÉ SESÉ

UNIVERSIDAD DE LA RIOJA
Cátedras Telefónica
UNIVERSIDAD COMPLUTENSE MADRID
UNIVERSITAT DE BARCELONA
UNIVERSITAT ROVIRA i VIRGILI
Logroño

Edited by

Mario Arias-Oliva

Jorge Pelegrín-Borondo

Kiyoshi Murata

Ana María Lara Palma

Manuel Ollé Sesé

ETHICOMP 2024

Smart Ethics in the Digital World

Proceedings of the ETHICOMP 2024

21th International Conference on the Ethical and Social Impacts of ICT

Logroño, Spain, March 2024

UNIVERSIDAD DE LA RIOJA

Cátedras Telefónica

UNIVERSIDAD COMPLUTENSE MADRID

UNIVERSITAT DE BARCELONA

UNIVERSITAT ROVIRA i VIRGILI

Logroño

PROCEEDINGS OF THE ETHICOMP* 2024

21th International Conference on the Ethical and Social Impacts of ICT

Logroño, La Rioja, Spain

March 13 – 15

Title	Smart Ethics in the Digital World
Edited by	Mario Arias-Oliva (Complutense University of Madrid), Jorge Pelegrín-Borondo (University of La Rioja), Kiyoshi Murata (Meiji University), Ana María Lara Palma (University of Burgos), Manuel Ollé Sesé (Complutense University of Madrid)
ISBN	978-84-09-58160-3
Local	Logroño, Spain
Date	March 1, 2024
Publisher	Universidad de La Rioja

Publisher: Universidad de La Rioja, www.unirioja.es

Cover designed by Universidad de La Rioja, Servicio de Comunicación, and Antonio Pérez-Portabella.

ISBN 978-84-09-58160-3

Presidency of the Scientific Committee

Mario Arias-Oliva, Complutense University of Madrid, Spain

Jorge Pelegrín-Borondo, University of La Rioja, Spain

Kiyoshi Murata, Meiji University, Japan

Ana María Lara-Palma, Universidad de Burgos, Spain

Manuel Ollé Sesé, Complutense University of Madrid, Spain

Scientific Committee

Antonio Fernández Portillo, University of Extremadura, Spain

Antonio Marturano, University of Rome Tor Vergata, Italy

Ana María Lara-Palma, Universidad de Burgos, Spain

Alba García Milon, University of La Rioja, Spain

Alejandro Cataldo, Talca University, Chile

Alicia Blanco, Rey Juan Carlos University, Spain

Alicia Izquierdo-Yusta, Universidad de Burgos, Spain

Camilo Prado Román, Rey Juan Carlos University, Spain

Cristina Olarte Pascual, University of La Rioja, Spain

David Cordon Benito, Complutense University of Madrid, Spain

Don Gotternbarn, Professor Emeritus at East Tennessee State University,USA

Emma Juaneda Ayensa, University of La Rioja, Spain

Graciela Padilla Castillo, Universidad Complutense de Madrid, Spain

Guadalupe Manzano-García, University of La Rioja, Spain

Jan Strohschein, Technische Hochschule Köln, Germany

Jani Koskinen, University of Turku, Finland

Jesús García de Madariaga Miranda, Universidad Complutense de Madrid, Spain

Joaquín Sánchez Herrera, Complutense University of Madrid, Spain

Jorge Pelegrín-Borondo, University of La Rioja, Spain

Jorge de Andrés Sánchez, Universitat Rovira i Virgili, Spain

José Antonio Fraiz Brea, Universidad de Vigo, Spain

Juan Carlos Yañez Luna, Autonoma University of San Luis de Potosí, Mexico

Katleen Gabriels, Maastricht University, The Netherlands

Kiyoshi Murata, Meiji University, Japan

Luis Blanco Pascual, University of La Rioja, Spain

Luz María Marín Vinuesa, University of La Rioja, Spain

Maria del Pilar Martínez Ruiz, Castilla - La Mancha University, Spain

Mario Arias-Oliva, Complutense University of Madrid, Spain

Marty J. Wolf, Bemidji State University, USA

Natalia Medrano Sáez, University of La Rioja, Spain

Nehme Khawly, Notre Dame University, Lebanon

Nuno Silva, Luisada University, Portugal

Orlando Lima Rua, Politechnic of Porto, Portugal

Pablo Gutierrez Rodríguez, León University, León, Spain

Paula Requeijo Rey, Complutense University of Madrid, Spain

Pedro Cuesta Valiño, Universidad de Alcalá de Henares, Spain

Pedro Isidoro González Ramírez, Autonoma University of San Luis de Potosí, Mexico

Ramón Alberto Carrasco González, Complutense University of Madrid, Spain

Presidency of the Organizing Committee

Organizing Committee

ETHICOMP Steering Committee

SUPPORTED BY

University of La Rioja

Complutense University of Madrid

Universitat Rovira i Virgili

Universitat de Barcelona

Ayuntamiento de Logroño

Cátedra URV-UB Telefónica Smart-Cities

Centre for Computing and Social Responsibility, De Montfort University

Centre for Business Information Ethics, Meiji University

To those who care about the human side beyond technology

The ETHICOMP conference series was launched in 1995 by the Centre for Computing and Social Responsibility (CCSR). Professor Terry Bynum and Professor Simon Rogerson were the founders and joint directors. The purpose of this series is to provide an inclusive international forum for discussing the ethical and social issues associated with the development and application of Information and Communication Technology (ICT). Delegates and speakers from all continents have attended. Most of the leading researchers in computer ethics as well as new researchers and doctoral students have presented papers at the conferences. The conference series has been key in creating a truly international critical mass of scholars concerned with the ethical and social issues of ICT. The ETHICOMP name has become recognised and respected in the field of computer ethics.

ETHICOMP previous conferences:

ETHICOMP 1995 (De Montfort University, UK)

ETHICOMP 1996 (University of Salamanca, Spain)

ETHICOMP 1998 (Erasmus University, The Netherlands)

ETHICOMP 1999 (LUISS Guido Carli University, Italy)

ETHICOMP 2001 (Technical University of Gdansk, Poland)

ETHICOMP 2002 (Universidade Lusiada, Lisbon, Portugal)

ETHICOMP 2004 (University of the Aegean, Syros, Greece)

ETHICOMP 2005 (Linköping University, Sweden)

ETHICOMP 2007 (Meiji University, Tokyo, Japan)

ETHICOMP 2008 (University of Pavia, Italy)

ETHICOMP 2010 (Universitat Rovira i Virgili, Spain)

ETHICOMP 2011 (Sheffield Hallam University, UK)

ETHICOMP 2013 (University of Southern, Denmark)

ETHICOMP 2014 (Les Cordeliers, Paris)

ETHICOMP 2015 (De Montfort University, UK)

ETHICOMP 2017 (Università degli Studi di Torino, Italy)

ETHICOMP 2018 (SWPS University of Social Sciences and Humanities, Poland)

ETHICOMP 2020 (University of La Rioja, Spain)

ETHICOMP 2021 (University of La Rioja, Spain)

ETHICOMP 2022 (University of Turku, Finland)

Table of contents

1. Explainable Artificial Intelligence (XAI) and Ethics

Ramón Carrasco, Complutense University of Madrid; Orlando Lima Rua, Polytechnic of Porto; Jorge Pelegrín-Borondo, University of La Rioja; Pedro Cuesta Valiño, Universidad de Alcalá de Henares; Pablo Gutierrez Rodríguez, León University

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) AND ETHICAL DECISION-MAKING IN BUSINESS

Gabriel Marín Díaz, José Javier Galán Hernández, José Luis Galdón Salvador

Universidad Complutense de Madrid (Spain), Universidad Politécnica de Valencia (Spain)

gmarin03@ucm.es; josejgal@ucm.es; jogalsal@doe.upv.es

EXTENDED ABSTRACT

Introduction

The question of whether machines can think, proposed by Alan M. Turing (Turing, 1950), has become increasingly relevant in today's world. Artificial Intelligence (AI) has made remarkable progress, surpassing human capabilities in various tasks traditionally associated with intelligence and creativity. However, as AI becomes more pervasive in decision-making processes within businesses, concerns regarding transparency, interpretability, and ethics arise. This article explores the importance of Explainable Artificial Intelligence (XAI) in facilitating ethical decision-making within the business context.

The Rise of XAI and the Paradox of Unexplainable Algorithms

Advancements in AI algorithms, particularly "black box" algorithms, have enabled machines to make complex decisions without human intervention. While these algorithms can yield impressive results, their decision-making logic often remains opaque and incomprehensible to humans. This poses a paradox: decision-making increasingly relies on AI systems that we struggle to understand fully. This lack of transparency raises concerns about accountability, biases, and potential risks in business operations (Kliegr et al., 2021).

The Need for Interpretability and the Ethical Implications

Recognizing the need for interpretability in AI algorithms, the concept of XAI has emerged. XAI aims to provide transparency and understandability in the decision-making process, enabling humans to comprehend and validate AI-driven decisions. Within the business context, interpretability becomes crucial as it allows decision-makers to assess the fairness, accuracy, and ethical implications of AI-generated recommendations or actions. Furthermore, interpretability helps identify and address potential biases in algorithms (Vallor & Rewak, 2019).

Legislative and Regulatory Considerations

Some sectors, such as finance, have recognized the importance of interpretability and have implemented regulations that mandate explanations for algorithmic decisions. However, broader awareness and understanding of interpretability in society are still limited. It is imperative for businesses to proactively engage with regulators, industry experts, and policymakers to shape legislation that ensures transparency, fairness, and ethical AI practices.

Ethical guidelines, standards, and accountability frameworks should be established to govern AI-driven decision-making in businesses (Glauner, 2022).

Advances in XAI and Mitigating Algorithmic Biases

Research and technological advancements in XAI offer promising solutions for addressing the interpretability challenge. Techniques such as rule-based explanations, visualizations, and model-agnostic approaches enable stakeholders to understand how AI algorithms arrive at decisions. Additionally, interpretability can help identify and mitigate biases present in training data or algorithm design (Molnar, 2019).

Creating a Culture of Ethical Decision-Making

Incorporating XAI into business processes requires a cultural shift towards ethical decision-making. Organizations should prioritize transparency, accountability, and human oversight in AI-driven systems. Decision-makers must possess a comprehensive understanding of AI capabilities, limitations, and potential biases to ensure responsible use (Bibal et al., 2020).

Model-agnostic interpretability algorithms in the context of XAI

The Model-agnostic algorithms provide techniques that can provide explanations for the decision-making process of any machine learning model, regardless of its underlying architecture or complexity. These algorithms focus on understanding and interpreting the behaviour of AI systems without relying on specific knowledge about how the models are built (Lundberg & Lee, 2017). Here are a few examples of their use in business settings:

- Decision Support Systems (Ribeiro et al., 2016): Model-agnostic algorithms can assist decision-makers in understanding the factors that contribute to AI-driven decisions. By providing explanations for each prediction or recommendation, these algorithms enable business professionals to gain insights into the underlying rationale and factors influencing the outcomes. This helps decision-makers make more informed choices.
- Risk Assessment and Compliance (Goodman & Flaxman, 2017): In industries such as finance, insurance, and healthcare, regulatory requirements often demand transparent and explainable decision-making processes. Model-agnostic algorithms allow businesses to identify potential biases, discrimination, or errors in the AI systems' outputs. By understanding the variables and features that influence the decision-making process, organizations can ensure compliance with regulations and mitigate potential risks.
- Customer Experience and Personalization (Marín Díaz et al., 2022): Model-agnostic algorithms can aid businesses in understanding the preferences and behaviour of their customers. By providing explanations for recommendations or personalized offerings, these algorithms allow organizations to provide transparency and gain customer trust. Moreover, they can help identify instances where AI-driven personalization might lead to unintended consequences or biases, enabling businesses to refine their algorithms and ensure fair and ethical treatment of customers.

- Fraud Detection and Cybersecurity (Zhang et al., 2022): Model-agnostic techniques can assist in identifying patterns and anomalies in large datasets, aiding in fraud detection and cybersecurity efforts. By explaining the features that contribute to suspicious activities or potential threats, these algorithms enhance the ability to interpret and validate AI systems' outputs, increasing the accuracy and effectiveness of fraud detection mechanisms.
- Process Optimization and Resource Allocation (Lakkaraju et al., 2016): Model-agnostic algorithms can uncover insights into complex business processes, enabling organizations to identify inefficiencies and optimize resource allocation. By providing explanations for the decisions made by AI models, businesses can pinpoint areas for improvement, streamline operations, and allocate resources more effectively.

Conclusions

By integrating ethics into AI practices, businesses can foster trust, maintain a positive reputation, and ensure the long-term viability and benefits of AI technologies within their operations. Embracing ethical AI not only aligns with societal expectations but also creates a competitive advantage by demonstrating responsible leadership in the ever-evolving landscape of AI-driven business practices.

- Transparency and accountability are paramount. Businesses should prioritize transparency in their AI systems and algorithms, ensuring stakeholders have a clear understanding of how decisions are made. This transparency builds trust and enables accountability for the outcomes produced by AI technologies.
- Fairness and non-discrimination should be prioritized. Businesses must actively identify and mitigate biases in their AI-driven processes that may result in discriminatory outcomes. Regular audits and evaluations are necessary to ensure equal opportunities and treatment for all individuals.
- Privacy and data protection are essential. Businesses must handle customer data responsibly, obtaining informed consent, implementing robust security measures, and adhering to relevant data protection regulations. Respecting privacy rights is critical for maintaining trust with customers and stakeholders.
- A human-centred approach is vital. AI should be designed to enhance human capabilities and improve decision-making, rather than replacing human workers. Businesses should prioritize the well-being of their employees and ensure that AI systems augment their skills and productivity.
- Ethical procurement and supply chain practices are necessary. Businesses should assess the ethical implications of AI technologies throughout their supply chains, ensuring that vendors and partners adhere to ethical standards and guidelines.
- Continuous monitoring and improvement are key. Ethical considerations should be an ongoing process, with businesses regularly evaluating the impact of AI on society, addressing emerging ethical challenges, and continuously improving their AI systems.

KEYWORDS: Decision-making, business processes, XAI, Ethics, AI practices.

REFERENCES

Bibal, A., Lognoul, M., de Streel, A., & Frénay, B. (2020). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, *0123456789*. https://doi.org/10.1007/s10506-020-09270-4

Glauner, P. (2022). *An Assessment of the AI Regulation Proposed by the European Commission*. *April 2021*, 119–127. https://doi.org/10.1007/978-3-030-99838-7_7

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Kliegr, T., Bahník, Š., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, *295*. https://doi.org/10.1016/j.artint.2021.103458

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 1675–1684. https://doi.org/10.1145/2939672.2939874

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*(Section 2), 4766–4775.

Marín Díaz, G., Galán, J. J., & Carrasco, R. A. (2022). XAI for Churn Prediction in B2B Models: A Use Case in an Enterprise Software Company. *Mathematics*, *10*(20). https://doi.org/10.3390/math10203896

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book*, 247. https://christophm.github.io/interpretable-ml-book

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 97–101. https://doi.org/10.18653/v1/n16-3020

Turing, A. M. (1950). Computing Machinery and Intelligence Author (s): A . M . Turing Source : Mind , New Series , Vol . 59 , No . 236 (Oct ., 1950), pp . 433-460 Published by : Oxford University Press on behalf of the Mind Association Stable URL : http://www.jstor.org/sta. *Mind*, *59*(236), 433–460.

Vallor, S., & Rewak, W. J. (2019). *An Introduction to Data Ethics*. 63. https://www.scu.edu/media/ethics-center/technology-ethics/IntroToDataEthics.pdf%0Ahttps://www.accenture.com/t20160629T012639Z__w__/us-en/_acnmedia/PDF-24/Accenture-Universal-Principles-Data-Ethics.pdf#zoom=50

Zhang, Z., Hamadi, H. Al, Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access*, *10*(September), 93104–93139. https://doi.org/10.1109/ACCESS.2022.3204051

AI EXPLAINABILITY, TEMPORALITY, AND CIVIC VIRTUE

Wessel Reijers, Tobias Matzner, Suzana Alpsancar, Martina Philippi

Paderborn University (Germany)

Wessel.reijers@uni-paderborn.de; tobias.matzner@uni-paderborn.de; suzana.alpsancar@uni-paderborn.de; martina.philippi@uni-paderborn.de

EXTENDED ABSTRACT

The notion that artificial intelligence (AI) has to be explainable has become entrenched in the public discourse concerning the ethical impacts of this emerging technology (Mittelstadt et al., 2016). Most notably, the stated reason for this concern is the property of neural networks to function as 'black box' models (Pasquale, 2015) that nonetheless perform certain modalities of reasoning. That is to say, these models 'reason' from particular inputs, which may consist of characters, pixels, or digital information in other modalities, to particular outputs, without transparently disclosing the process of this reasoning. This is often contrasted with 'good old fashioned AI' (GOFAI) models that use decision trees which – in principle – can be followed by a human expert from input to output. The problem with neural nets, implemented in programs like ChatGPT and Dall-E, is that they can potentially influence or even autonomously make decisions about human affairs that cannot ex-post be explained by human interpreters – even if these are experts. At most, humans may figure out the particular artificial neurons that had an important influence on a decision.

Yet, the feasibility and relevance of the principle of explainability has been questioned. Robbins (2019) has argued that in fact, people are not required to explain every decision they make. Instead, explainability only becomes an issue in exceptional circumstances when the outcome of a particular decision requires explanation. It would therefore be unreasonable and unhelpful to insist on a standard for AI systems that does not apply to human decision-making. Moreover, meaningful human control over AI decision-making, which is arguably one of the aims of explainability, can be achieved by other means – for instance through proper legislation. Others have argued that explainability should not be reduced to explicability (i.e., accounting for the explanandum) but should involve the social context, considering it as a set of social practices (Rohlfing et al., 2021). Indeed, explaining takes place in a social context, and moreover has different modalities.

From this perspective, explainability as such is neither a mere technical matter, nor is it in any case relevant, nor is it a singular phenomenon. This paper proposes an initial way to grapple with these difficulties, by considering – first of all – the role of temporality in different modalities of explaining, and – secondly – the normative perspective of civic virtue to evaluate these different modalities, which then raises distinct requirements for explainability given distinct social contexts.

Let us start with the consideration of temporality, as it offers a ground to consider different modalities of explanation. In the *Rhetoric*, Aristotle set out the idea that argumentation occurs in different temporal modalities. It can be past-oriented, in which case it is *forensic*, explaining what has happened by reference to memory and traces. It can be present-oriented, in which case it is *epideictic*, explaining why a person or act deserves blame or honor, or the assignment

of virtue or vice. It can, furthermore, be future-oriented, in which case it is *deliberative*, explaining why particular future outcomes should or should not be supported. AI systems can, in principle, be involved in all three of these modalities of explaining, but they confront us with different normative requirements when they do. Forensic explanations, for instance, put forward requirements concerning historical proof, whereas deliberative explanations put forward requirements concerning (political) vision and conviction.

To make sense of these normative requirements, we may also draw from Aristotle. For in Aristotle, as Johnstone argues, (2023), ethics, rhetoric, and politics are fundamentally interrelated. Modalities of explanation, in other words, have a bearing on ethical and political life, in that they affect human virtues. Virtue is therefore a valid point of departure, as Vallor has forcefully argued (2016) in the context of technology ethics, in considering how AI affects explainability in a normative sense. Yet, virtue is also primarily grounded in the life of the individual, being anchored in *eudaimonia*, and does not yet offer the resources to bridge the gap between the ethics of the individual and the politics of the community. Civic virtue, developed in Aristotle's *Politics*, does offer this transitory concept, for it always mediates between the aim of the individual and the aim of the political community. As such, it is also inherently concerned with technology, as the technological infrastructure is a primary concern of the mode by which civic virtue is cultivated and enacted.

Strikingly, the distinct modalities of explanation and the distinct notions of civic virtue in political philosophy can each be grounded in a consideration of temporality. Like modalities of explanation, civic virtue can be past-, present-, and future-oriented. Past-oriented civic virtue finds its most vocal adherents in liberal and neo-republican thought, where it in an instrumental quality that draws from a history of reputational events, cultivating a sense of civility amongst a population (Pettit, 1997). Present-oriented civic virtue finds its footing in classical republican thought, where it requires institutional structures for the support of practices that aim at internal goods (MacIntyre, 2007). Future-oriented civic virtue finds its basis in existential republican thought, which puts forward the requirement of a durable public sphere that supports political action in concert (Arendt, 1958).

How do these different modalities of civic virtue help us to think through the modalities of explainable AI? First, they help us to consider the plurality of explanations insofar as they relate to different modalities of civic virtue. To give an example: when faced with a reputation-building AI (e.g., a credit scoring mechanism), the aim of such a system is to mediate past-oriented civic virtue; in that reputation building implies a historical record of reputational events. Such a mode of civic virtue put forward requirements deriving from forensic explanations. In other words, for such an AI to cultivate rather than to corrupt civic virtue, its explainability would need to safeguard requirements of – amongst others – historical proof. When faced with a more explicitly political AI (e.g., the use of AI in mass online deliberation), the aim of such a system is to mediate future-oriented civic virtue; in that it supports deliberative decision-making about alternative political pathways. Such a mode of civic virtue puts forward requirements deriving from deliberative explanations. Differently put, for such an AI to cultivate rather than to corrupt civic virtue, its explainability would need to respect requirements of – amongst others – political conviction. It goes without saying that the latter requirements would be rather more stringent and putting up a higher bar than the former.

What this tells us is, foremost, that not every explanation is equal. Whether an explanation is required at all, and what modality it should be in, depends on the temporal mode of the human

activities that an AI system affects. In a shorthand manner, one could argue that the more AI infringes onto the political realm, the more stringent explainability requirements will be. At the same time, the modality of those requirements will also change, for instance shifting from forensic to deliberative requirements.

KEYWORDS: Explainability, AI, civic virtue, temporality.

REFERENCES

Arendt, H. (1958). *The Human Condition* (Vol. 24, Issue 1). University of Chicago Press. https://doi.org/10.2307/2089589

Johnstone, C. L. (2023). *An Aristotelian Trilogy: Ethics, Rhetoric, Politics, and the Search for Moral Truth*.

MacIntyre, A. (2007). *After Virtue: A Study in Moral Theory*. University of Notre Dame Press.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Pasquale, F. (2015). *The Black Box Society*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Pettit, P. (1997). *Republicanism: A Theory of Freedom and Government*. Oxford University Press.

Robbins, S. (2019). A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*, *29*(4), 495–514. https://doi.org/10.1007/s11023-019-09509-3

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Hab-Umbach, R., Horwath, I., Hullermeier, E., Kern, F., Kopp, S., Thommes, K., Ngonga Ngomo, A.-C., Schulte, C., Wachsmuth, H., Wagner, P., & Wrede, B. (2021). Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(3), 717–728. https://doi.org/10.1109/TCDS.2020.3044366

Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.

ETHICS UNVEILED: ILLUMINATING THE PATH OF AI INTEGRATION IN HIGHER EDUCATION

Jorge Pelegrín-Borondo, Cristina Olarte-Pascual, Luis Blanco-Pascual, Alba García-Milon

University of La Rioja (Spain)

jorge.pelegrin@unirioja.es; cristina.olarte@unirioja.es; luis.blanco@unirioja.es; alba.garciam@unirioja.es

EXTENDED ABSTRACT

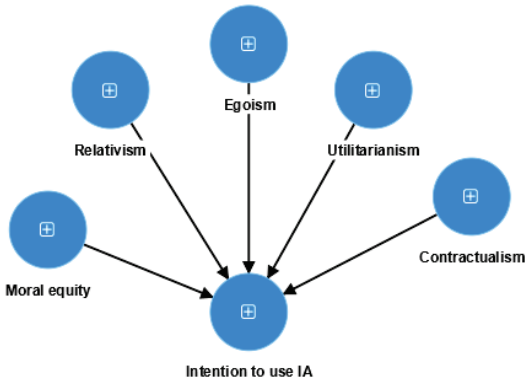
In the field of higher education, Artificial Intelligence (AI) presents both new possibilities and challenges (Silander & Stigmar, 2019). It offers opportunities to improve governance effectiveness and efficiency, benefiting students, teachers, administrative staff, and researchers (Nasrallah, 2014). Therefore, there is a need for integrating AI into higher education (Stefan & Sharon, 2017). However, the use of AI-based technologies for teaching and learning raises ethical issues (Celik, 2023). AI tools can exhibit systematic errors, leading to discrimination against students from diverse backgrounds and compromising inclusiveness in education (De Cremer & De Schutter, 2021; Dietvorst et al., 2018). Other ethical concerns associated with AI include content moderation, environmental impact, and the risk of copyright infringement (Cooper, 2023).

Currently, teachers face the dilemma of whether to encourage or discourage students from using AI. In this decision, teachers' ethical considerations regarding their students' use of this technology can be crucial in determining their role as integrators or opponents of AI. Ethics allows addressing the controversy between the potential benefits of technological progress and the duty not to jeopardize that progress (Olarte-Pascual, Pelegrín-Borondo, Reinares-Lara, Arias-Oliva, 2021). However, the impact of different dimensions of ethical judgment on this decision remains unexplored. This research aims to address this question, focusing on the widely recognized AI platform ChatGPT, which has gained global attention and public interest. Recent news in Spain indicates that university students are extensively using ChatGPT (Planas Bou, 2023).

Reidenbach and Robin (1990) developed the Multidimensional Ethical Scale (MES), which proposes that individuals use multiple reasons to make ethical judgments. Originally consisting of eight items measuring three subscales, the MES was distilled and validated from an initial inventory of 33 items (Reidenbach & Robin, 1990, p. 639). The MES (1990) and its modified versions (e.g. Kadić-Maglajlić et al., 2017; Mudrack & Mason, 2013; Pelegrín-Borondo et al., 2020) have been widely used to explain the influence of ethical judgment on behavior. Shawver and Sennetti (2009) proposed the Composite MES, a modification that incorporates items from the five major normative ethical theories. The Composite MES has been extensively used to explain the impact of ethical judgments on behavior (e.g., Kara et al., 2016; Manly et al., 2015; Mudrack & Mason, 2013). It includes the dimensions of moral equity, relativism, utilitarianism, egoism, and contractualism (Nguyen & Biderman, 2008; Reidenbach & Robin, 1990).

Building upon this theoretical framework, the authors propose to investigate how the different dimensions of ethical judgment influence university professors' intention to encourage their students to use AI in their tasks and academic activities. To achieve this, the following model is proposed (Figure 1).

Figure 1. Proposed model



A self-administered survey was conducted among university professors from Business Faculties in Spain to test the proposed model. An invitation was sent to all professors through a national association representing business faculties. A total of 270 valid surveys were collected, with 53% males and 47% females. The average age was 49.95 years (SD = 9.88). The MES Composite scale by Shawver and Sennetti (2009) was used to measure ethical judgment dimensions, employing an 11-point semantic differential scale. The professors' intention to encourage their students to use AI in academic activities was measured using a 2-item Likert scale based on Venkatesh and Davis's (2000) Technology Acceptance Model TAM2. The statistical analysis of the model was conducted using PLS (Partial Least Squares).

Regarding the results, the reliability and validity of the scales were examined. One item from the relativism dimension was removed due to convergent validity issues. The final scales demonstrated satisfactory reliability, convergent validity, and discriminant validity, as shown in Table 1.

Table 1. Composite reliability, Cronbach's alpha, AVE (convergent validity) and discriminant validity.

Construct	**Composite reliability > 0.7**	**Cronbach's *Alpha* > 0.7**	**AVE > 0.5**	**HTMT**				
				ME	**R**	**E**	**U**	**C**
Moral Equity (ME)	0.969	0.970	0.942					
Relativism (R)	0.863	0.863	0.880	0.898				
Egoism (E)	0.938	0.938	0.941	0.857	0.853			
Utilitarianism (U)	0.859	0.870	0.876	0.835	0.866	0.884		
Contractualism (C)	0.965	0.965	0.966	0.830	0.852	0.777	0.835	
Intention to use (IU)	0.958	0.958	0.960	0.764	0.707	0.743	0.669	0.699

Table 2 displays the values of R^2 and Q^2, the path coefficients (direct effects), and p-values for each antecedent variable of professors' intention for their students to use AI. The R^2 for the model of AI use intention was high ($R^2 = 0.629$), and the Q^2 provided by PLS Predict was greater than 0.5 ($Q^2 = 0.565$). This indicates that the dimensions of ethical judgment have explanatory

and predictive power over professors' intention for their students to use AI. In Table 2, it is shown that the dimensions of moral equity, egoism, and contractualism positively influence the intention to use AI.

Table 2. Effect on the endogenous variables.

	R^2	Q^2	Path coefficient	*p*-value
INTENTION TO USE AI	0.585	0.565		
Moral Equity =>(+) Intention to use IA			0.401	0.000
Relativism =>(+) Intention to use IA			-0.014	0.850
Egoism =>(+) Intention to use IA			0.304	0.001
Utilitarianism =>(+) Intention to use IA			-0.076	0.310
Contractualism =>(+) Intention to use IA			0.195	0.013

The findings show that professors' ethical judgment dimensions have a differentiated impact on their intention to promote student use of AI in tasks and teaching activities. Three dimensions, namely moral equity, egoism, and contractualism, positively influence this intention. Among them, moral equity has the strongest explanatory power, indicating that perceiving AI use as fair motivates teachers to encourage it. Egoism is the second influential dimension, suggesting that personal benefits from student AI use increase teachers' inclination to promote it. Contractualism is the third influencing dimension, indicating that perceiving an implicit agreement within the university for AI use leads to greater encouragement. However, no evidence supports the impact of relativism and utilitarianism dimensions on professors' intention to promote student AI use. These conclusions emphasize the significance of considering professors' ethical perceptions when integrating AI in education and provide valuable insights for developing effective strategies for AI integration in teaching.

KEYWORDS: Artificial intelligence, ethical concerns, higher education, intention to use.

ACKNOWLEDGEMENTS: The authors gratefully acknowledge the support from the University of La Rioja for funding Teaching Innovation Projects during the academic year 2023-24, as well as the grant given to the COBEMADE Research Group at the University of La Rioja.

REFERENCES

Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. Computers in Human Behavior, 138, 107468.

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. Journal of Science Education and Technology, 32(3), 444-452.

De Cremer, D., & De Schutter, L. (2021). How to use algorithmic decision-making to promote inclusiveness in organizations. *AI and Ethics*, 1(4), 563–567. https://doi. org/10.1007/s43681-021-00073-0

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Kadi´c-Maglajli´c, S., Arslanagi´c-Kalajd˘zi´c, M., Micevski, M., Michaelidou, N., & Nemkova, E. (2017). Controversial advert perceptions in SNS advertising: The role of ethical judgement and religious commitment. *Journal of Business Ethics*, 141(2), 249–265. https://doi.org/10.1007/s1055.1-015-2755-5

Kara, A., Rojas-M´endez, J. I., & Turan, M. (2016). Ethical evaluations of business students in an emerging market: Effects of ethical sensitivity, cultural values, personality, and religiosity. *Journal of Academic Ethics*, 14(4), 297–325. https://doi. org/10.1007/s10805-016-9263-9

Manly, T. S., Leonard, L. N., & Riemenschneider, C. K. (2015). Academic integrity in the information age: Virtues of respect and responsibility. *Journal of Business Ethics*, 127 (3), 579–590.

Mudrack, P. E., & Mason, E. S. (2013). Ethical judgments: What do we know, where do we go? Journal of Business Ethics, 115(3), 575–597. https://doi.org/10.1007/s10551-012-1426-z

Nasrallah, R. (2014). Learning outcomes role in higher education teaching. *Education, Business and Society*, 7(4), 257–276. https://doi.org/10.1108/EBS-03-2014-0016

Nguyen, N. T., & Biderman, M. D. (2008). Studying ethical judgments and behavioral intentions using structural equations: Evidence from the multidimensional ethics scale. *Journal of Business Ethics,* 83(4), 627–640. https://doi.org/10.1007/s10551-007-9644-5

Olarte-Pascual, C., Pelegrín-Borondo, J., Reinares-Lara, E. Arias-Oliva, M. (2021). From wearable to insideable: Is ethical judgment key to the acceptance of human capacity-enhancing intelligent technologies? *Computers in Human Behavior*, 114, 106559.

Pelegrín-Borondo, J., Arias-Oliva, M., Murata, K., & Souto-Romero, M. (2020). Does ethical judgment determine the decision to become a cyborg? *Journal of Business Ethics*, 161(1), 5–17. https://doi.org/10.1007/s10551-018-3970-7

Planas Bou, C (2023). Universitarios y adolescentes se pasan en masa a ChatGPT para hacer trabajos (y exámenes). El Periódico (9-05-2023). https://www.elperiodico.com/es/sociedad/20230508/chatgpt-universidad-escuelas-inteligencia-artificial-estudiantes-deberes-examenes-86837251

Reidenbach, R. E., & Robin, D. P. (1988). Some initial steps toward improving the measurement of ethical evaluations of marketing activities. *Journal of Business Ethics*, 7, 871–879. https://doi.org/10.1007/BF00383050

Reidenbach, R. E., & Robin, D. P. (1990). Toward the development of a multidimensional scale for improving evaluations of business ethics. *Journal of Business Ethics*, 9(8), 639–653. https://doi.org/10.1007/BF00383391

Shawver, T. J., & Sennetti, J. T. (2009). Measuring ethical sensitivity and evaluation. *Journal of Business Ethics*, 88(4), 663–678. https://doi.org/10.1007/s10551-008-9973-z

Silander, C., & Stigmar, M. (2019). Individual growth or institutional development? Ideological perspectives on motives behind Swedish higher education teacher training. *Higher Education: The International Journal of Higher Education Research*, 77, 265–281. https://doi.org/10.1007/s10734-018-0272-z

Stefan, A. D. P., & Sharon, K. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 1, 3–13. https://doi.org/10.1186/s41039-017-0062-8

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science*, 46, 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

UNPACKING THE PURPOSES OF EXPLAINABLE AI

Suzana Alpsancar, Tobias Matzner, Martina Philippi

Paderborn University (Germany)

suzana.alpsancar@uni-paderborn.de; tobias.matzner@uni-paderborn.de,
martina.philippi@uni-paderborn.de

EXTENDED ABSTRACT

AI systems are being increasingly deployed in various societal fields. Much of the AI technology that is contributing to this success, particularly machine learning (ML), is opaque, meaning that how it works or why it exhibits a particular behavior or performance is not (immediately) obvious for a variety of reasons (Burrell 2016). Numerous cases have shown that this opacity can become problematic: Some ML models have been easy to trick and have exhibited Clever-Hans effects, domain shifts, and overfitting. Others have incorrectly influenced grave decisions such as the probability of death in a patient with pneumonia (Cabitza, Rasoini, and Gensini 2017), or have been subject to adversarial attacks (Gilpin et al. 2018). Scandals and debates about the biases and fairness of, for example, COMPAS recidivism prediction software (Angwin et al. 2016) have contributed to growing ethical concerns about AI. According to the literature review by Tsamados et al. (2022), these ethical concerns can be distinguished into two normative concerns (unfair outcomes and transformative effects), three epistemic concerns (inconclusive evidence, inscrutable evidence, and misguided evidence), as well as the concern of traceability (the possibility of tracing the chain of events of factors that brought about a given outcome) that affects all other concerns. Whereas the normative concerns relate explicitly to ethical impacts such as unintended consequences or biases of AI systems, the epistemic concerns relate to the justifiability of the outcome of AI systems, and this, in turn, may evoke morally critical decisions.

For all these reasons, it would be game changing if both adopters (e.g., medical practitioners) and affected individuals (e.g., patients) would be able to adequately assess the performance and limitations of AI systems. There is a widespread at least implicit assumption in the field that ”explainability is a suitable means for facilitating trust in a stakeholder”, what Kästner et al. (2021) have depicted as the ”Explainability-Trust-Hypotheses”. Against this background, explainable AI (xAI) has become highly valorized. Explainability is considered as necessary for robust and trustworthy AI applications and, hence, for their commercial success (Arya et al. 2019). As several meta-reviews have shown (Hagendorff 2020; Jobin, Ienca, and Vayena 2019; Morley et al. 2020), explainability is a central element of all voluntary commitments and ethical guidelines for AI in industry, research, and policymaking. For instance, the European Commission’s High Level Expert Group on Artificial Intelligence literally links the research field of xAI to its agenda of building trustworthy AI:

> For a system to be trustworthy, we must be able to understand why it behaved a certain way and why it provided a given interpretation. A whole field of research, Explainable AI (xAI) tries to address this issue to better understand the system’s underlying mechanism and final solutions. (High Level Expert Group 2019, 21)

The latest regulatory requirements echo this valorization of explainability, especially within the EU where legislation might expand existing laws into a right to explanation (Wachter, Mittelstadt, and Floridi 2017) and where the proposed AI Act sets new obligations to ensure transparency, which is often directly linked to explainability (EC 2021).

From a philosophical perspective, the call for xAI rests on a normative claim: "good AI is xAI" or even the stronger claim "only xAI is good AI." This valorization runs the risk of being overgeneralized because explanations are not per se useful, appropriate, or demanded. Clearly, the practical use of xAI depends on whether the explanation is needed at all, whether it is appropriate for the explainees, and whether it is understandable. Previous literature reveals some voices that are critical of the value of explaining. For instance, Robbins argues that the principle of explicability[1] is misdirected. He points out three misgivings: (1) It is not the process of coming up with a decision, but the decision (or action) itself that is in need of an explanation. (2) It makes no sense to demand from all AI systems that they should explain themselves, because there are many applications with a low risk (in terms of potential harm of moral weight). (3) For high-risk applications, it is contradictory to demand explicability from the AI system, because they are designed precisely to serve areas in which we do not know what parameters to consider (Robbins 2019, 509).

We agree with Robbins' basic intervention that not all AI systems must necessarily be explainable, that explainability is not a value in itself, and that explainability is not always useful. However, we disagree with his classificatory theoretical perspective: Neither algorithms nor decisions can be classified per se as needing or not needing explanations—which is what he suggests as being a better strategy. Instead, we follow a practice theoretical approach in arguing that explainability should neither be conceptualized as a trait of a technical artifact nor as a property of a mere decision or an act, but as a disposition of a given sociotechnical system that must be materialized in practices of explaining within given socially structured contexts.

If we account for explainability as an instrumental value, we need to explicate what explainability is meant to deliver from both an ethical perspective and the perspective of respective users (or other stakeholders). Hence, we need to answer the following normative questions when it comes to adequately evaluating the goodness of explanations:

1. When is an explanation ethically obligatory?
2. When is an explanation individually helpful (to whom for which purpose)?
3. What characterizes a good explanation (in light of 1. and 2.)?

Currently, these rarely explicated questions are usually answered by referring to those motives that give reasons for developing xAI in the first place—that is, naming what xAI is meant to be good for. These for-the-sake-of relations can be systematized into three categories:

a. Functional purposes such as keeping a system running, debugging it, or improving it technically (developing AI)

[1] Robbins adopts the language of Floridi et al. (Floridi et al. 2018) and argues against the claim that all AI must be explicable in their sense, that is of guaranteeing "meaningful human control." His objections, however, can be related to a generalization of explainability, not just to its utility for this interpretation of "ethical assurance."

b. Social or economic purposes such as satisfying so-called users' needs, e.g., explaining apparently awkward social robot behavior (deploying AI)

c. Normative purposes, i.e. respecting ethical values and principles or meeting legal requirements, e.g., presenting reasons for loan rejections to render the decision-making process contestable (governing AI)

The first type of purposes echoes the experiences of those who develop and optimize ML components, e.g., the first techniques for explaining AI had been developed by ML experts for other experts, e.g., in the context of the Pascal Visual Object Classes (VOC) Challenge, which serves as a benchmark for object recognition/detection in ML, to unmask Clever-Hans effects (Everingham et al. 2015). With the wider distribution of AI systems (AIS) in various societal fields, the xAI community increasingly draws attention to lay persons (users, operators, domain experts) and to meet ethical and legal demands. Here, there is a strong motivation to mimic interpersonal interaction. For instance, de Graaf and Malle (2017) argue that the entire interaction with nonhuman agents, including explanations, should correspond to the user's expectations, namely their underlying intentional framework. If AI systems do not reveal their intention, users find them "unsettling and creepy" (de Graaf and Malle 2017, 20).

In terms of the ethical demands, much has been said about the challenges of moving 'from principles to practice' (Rességuier and Rodrigues 2020). Very little has been discussed about the conditions under which certain purposes can be considered adequate: When is it necessary, helpful, or adequate for an xAI system to serve the purpose of particular ethical principles, and how does this relate to other purposes xAI is meant to serve?

In our paper we aim to put the goodness of the presumed purposes, xAI is meant to serve, into question and we want to particularly question how functional, economic, and ethical purposes relate to each other. As we think that such an evaluation only makes sense in a contextualized setting, we will pursue our analyses by comparing two stylized use cases: deploying automated and connected vehicles and deploying algorithmic decision-making systems in a healthcare facility.

KEYWORDS: Explainable AI, valorization of xAI, purposes, ethical demands, users' needs.

REFERENCES

Angwin, Julia, Larson, Jeff, Mattu, Surya and Kirchner, Lauren Kirchner (2016). "Machine Bias. There's Software used across the country to predict future criminals. And it's biased against blacks." ProPublica, accessed March 27, 2022. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arya, Vijay, Bellamy, Rachel K. E., Chen, Pin-Yu, Dhurandhar, Amit, Hind, Michael, Hoffmann, Samuel C., Houde, Stephanie, et al (2019). "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques." CoRR abs/1909.03012. arXiv: 1909.03012

Burrell, Jenna (2016). "How the machine thinks: Understanding opacity in machine learning algorithms." Big Data & Society 3 (1): 1–12. https://doi. org/10.1177/2053951715622512

Cabitza, Federico, Rasoini, Raffaele, and Gensini, Gian Franco (2017). "Unintended Consequences of Machine Learning in Medicine." JAMA 318 (6): 517–518. https://doi.org/10.1001/jama.2017.7797

De Graaf, Maartje MA, and Malle, Bertram F. (2017). "How people explain action (and autonomous intelligent systems should too)." In 2017 AAAI Fall Symposium Series.

EC (2021). "Proposal for regulation of the European parliament and of the council Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts." European Commission. Digital Strategy, accessed September 12, 2022. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

Everingham, Mark, Eslami, SM Ali, Van Gool, Luc, Williams, Christopher KI, Winn, John and Zisserman Andrew (2015). "The pascal visual object classes challenge: A retrospective." International journal of computer vision 111: 98–136. https://doi.org/10.1007/s11263-014-0733-5.

Floridi, Luciano, Cowls, Josh, Beltrametti, Monica, Chatila, Raja, Chazerand, Patrice, Dignum, Virginia, Luetge, Christoph et al (2018). "AI4People–An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." Minds and Machines 28 (4): 689–707. https://doi. org/10.1007/s11023-018-9482-5.

Gilpin, Leilani H, Bau, David, Yuan, Ben Z, Bajwa, Ayesha, Specter, Michael and Kagal, Lalana (2018). "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning." In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics.

Hagendorff, Thilo (2020). "The Ethics of AI Ethics: An Evaluation of Guidelines." Minds and Machines 30 (1): 99–120. https://doi.org/10.1007/s11023-020-09517-8

High Level Expert Group (2019). "Ethics Guidelines for Trustworthy AI." European Commission, accessed October 29, 2021. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Jobin, Anna, Ienca, Marcello and Vayena, Effy (2019). "Artificial Intelligence: the global landscape of ethics guidelines." Nat. Mach. Intell., 389–399. https: //doi.org/10.1038/s42256-019-0088-2

Kästner, Lena, Langer, Markus, Lazar, Veronika, Schomäcker, Astrid, Speith, Timo and Sterz, Sarah (2021). "On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness." In 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW), 169–175. https://doi.org/10.1109/REW53955.2021.00031

Morley, Jessica, Floridi, Luciano, Kinsey, Libby and Elhalal, Anat (2020). "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 26:2141–2168. https://doi.org/10.2139/ssrn.3830348

Rességuier, Anais, and Rodrigues, Rowena (2020). "AI ethics should not remain toothless! A call to Bring back the teeth of ethics." Big Data & Society 7 (2): 1–5. https://doi.org/10.1177/2053951720942541

Robbins, Scott (2019). “A Misdirected Principle with a Catch: Explicability for AI.” Minds and Machines 29 (4): 495–514. https://doi.org/10.1007/s11023-019-09509-3

Tsamados, Andreas, Aggarwal, Nikita, Cowls, Josh, Morley, Jessica, Roberts, Huw, Taddeo, Mariarosaria and Floridi, Luciano (2022). “The ethics of algorithms: key problems and solutions.” AI & SOCIETY 37 (1): 215–230. https://doi.org/10.1007/s00146-021-01154-8

Wachter, Sandra, Mittelstadt, Brent and Floridi, Luciano (2017). “Why a right to explanation of automated decision-making does not exist in the general data protection regulation.” International Data Privacy Law 7 (2): 76–99.

ARTIFICIAL INTELLIGENCE IN SCIENCE: SHUT UP AND CALCULATE

Ramón Alberto Carrasco, Mario Arias-Oliva, Jesús Serrano-Guerrero, Francisco Chiclana

Universidad Complutense De Madrid (Spain), Universidad Complutense De Madrid (Spain), Universidad Castilla-La Mancha (Spain), De Montfort University Leicester (United Kingdom)

ramoncar@ucm.es; mario.arias@ucm.es; jesus.serrano@uclm.es; chiclana@dmu.ac.uk

EXTENDED ABSTRACT

Quantum mechanics is astonishing, both in terms of its accuracy and its interpretation. Not even geniuses like Albert Einstein could imagine the underlying reality within this theory, which, on the other hand, has formulations that are fulfilled with very high precision. In order to reach a consensus on its interpretation, the best physicists came together between 1925 and 1927, giving rise the so-called "Copenhagen interpretation". However, for many authors, this interpretation implies a refusal to grasp the truth. Thus, David Mermin (1989) coined the phrase "shut up and calculate!" to summarize this Copenhagen interpretation or, rather, attitude. This sentence summarizes the approach of many scientists who apply this accuracy theory mechanically, without considering anything else regarding its interpretation.

On the other hand, we are living a boom of the artificial intelligence (AI) that lives its "golden spring". This is mainly due to the advances of the machine learning algorithms, which learn from vast amounts of data, produce at very high velocity, in various structured and non-structured forms (including audios, videos, images, natural language, etc.). Among these algorithms, the ones that do the best job are the so-called black box algorithms, which are not interpretable by humans, including deep learning based on artificial neural networks (Das et al., 2020). Therefore, the situation we are facing involve the delegation of decision-making in favour of AI, even although these decisions are based on algorithms that are non-compressible for us (Marín & Carrasco, 2021). The use of these black-box algorithms is spreading to all fields: military, healthcare, education, finance, business, marketing, science, etc.

Indeed, the scientist world is not an exception and, more and more this kind of algorithms are being used in researches. The reflection is as follows: Are scientists in general, particularly those not specialized in AI, aware of the non-interpretability of the outcomes generated by this type of AI?

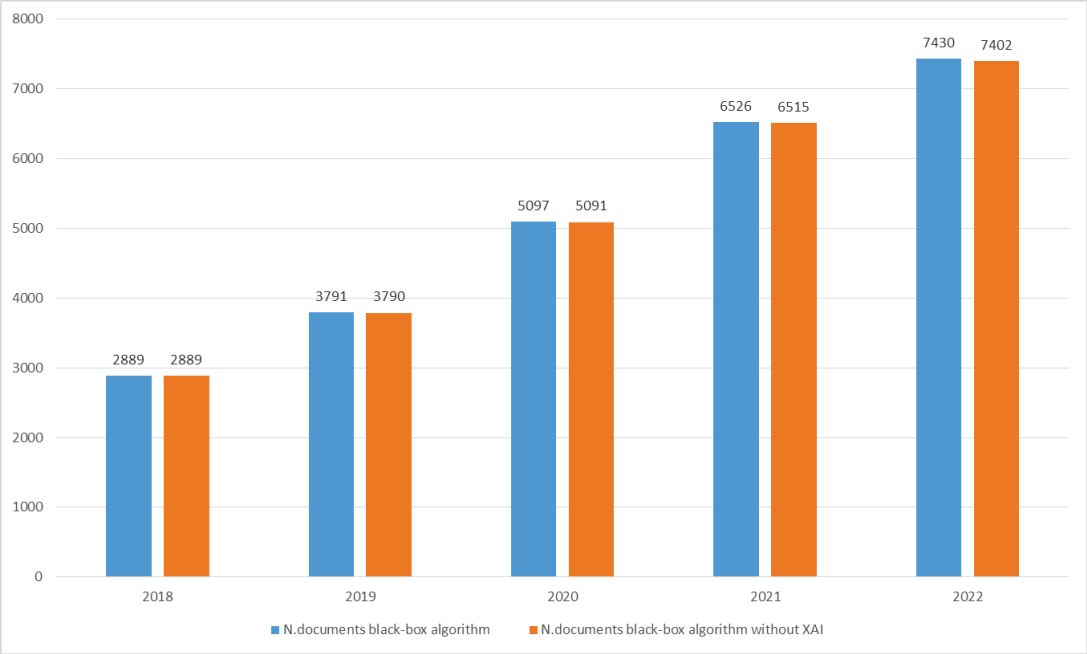
The field of explainable artificial intelligence (XAI) emerged with the aim of making the results of these black-box algorithms more interpretable (Monje et al., 2022; Marín et al., 2022; Gunning & Aha, 2019). So, we could rephrase the previous question: Is the scientific community employing XAI to enhance the interpretability of black-box algorithms?

To help clarify this question, we are relying on the documents published in the Web of Science (WoS) database. While exploring just applications of black-box algorithms, we conduct a query exclusively within the social science domains in the core collection of WoS. In other words, this approach aims to exclude studies within the AI field itself. We are going to conduct the exploration through two paths. Firstly, let us obtain documents that use black-box algorithms in social sciences. In the second place, we add that they do not use XAI techniques. Both queries are displayed in Table 1 and the results are showed in the Figure 1.

Table 1. WOS queries description (*year* between 2018 and 2022).

Query description	**Query**
Documents using of black-box algorithms in social sciences	TS = (Neural Networks OR Deep Learning OR Support Vector Machine* OR Random Forests OR Gradient Boosting Machine* OR Extreme Learning Machine* OR Kernel Machin* OR Kernel-based Learning Machine* OR Long Short-Term Memory OR Ensemble Classification Models OR Emsemble model) AND PY=(*year*) AND WC=("Agricultural Economics & Policy" OR Anthropology OR "Area Studies" OR Art OR "Asian Studies" OR "Behavioral Sciences" OR Business OR "Business, Finance" OR Classics OR "Construction & Building Technology" OR "Criminology & Penology" OR "Cultural Studies" OR Dance OR Demography OR "Development Studies" OR Economics OR "Education & Educational Research" OR "Education, Scientific Disciplines" OR "Education, Special" OR "Environmental Studies" OR "Ethnic Studies" OR Folklore OR Geography OR History OR "Humanities, Multidisciplinary" OR "Information Science & Library Science" OR "International Relations")
Documents using of black-box algorithms in social sciences without XAI	TS=(Neural Networks OR Deep Learning OR Support Vector Machine* OR Random Forests OR Gradient Boosting Machine* OR Extreme Learning Machine* OR Kernel Machin* OR Kernel-based Learning Machine* OR Long Short-Term Memory OR Ensemble Classification Models OR Emsemble model) NOT TS = (XAI OR “Explainable Artificial Intelligence”) AND PY=(*year*) AND WC=("Agricultural Economics & Policy" OR Anthropology OR "Area Studies" OR Art OR "Asian Studies" OR "Behavioral Sciences" OR Business OR "Business, Finance" OR Classics OR "Construction & Building Technology" OR "Criminology & Penology" OR "Cultural Studies" OR Dance OR Demography OR "Development Studies" OR Economics OR "Education & Educational Research" OR "Education, Scientific Disciplines" OR "Education, Special" OR "Environmental Studies" OR "Ethnic Studies" OR Folklore OR Geography OR History OR "Humanities, Multidisciplinary" OR "Information Science & Library Science" OR "International Relations")

Figure 1. Documents using of black-box algorithms in social sciences.



Source: self-elaboration based on WOS (2023)

These results indicate that most of the research in social science using black-box AI does not consider the interpretability of this AI. The primary purpose of utilizing this AI is predicting certain variables. However, few inquire about the interpretation of these black models. In addition, as expected, it is possible to deduce from Figure 1 that the usage of this type of AI is increasing. Many scientists are convinced of the effectiveness of these algorithms primarily due to the accuracy of the predictions but they do not worry excessively about the interpretation. Once again in the history of science, we are moving towards the "shut up and calculate!" approach.

We must emphasize that the non-interpretability of AI could represent a neglect of another important aspect related to AI research: responsibility, bias, fairness, ethical considerations, etc (Giovanola & Tiribelli, 2023; Yapo & Weiss, 2018). To the extent that these aspects are essential, in many cases, within the field of science, we must be demanding in our endeavor to comprehend this black AI. Put more pragmatically, in the utilization of XAI techniques. While these techniques are not flawless, they at least assist us in approaching the interpretation of the previously mentioned aspects.

KEYWORDS: XAI, black-box algorithms, interpretable AI, ethical AI.

ACKNOWLEDGMENTS: The co-author Ramón Alberto Carrasco González was funded by the Madrid Government (Comunidad de Madrid-Spain) under the Multi-annual 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).

This work was also supported by the Project PID2019-103880RB-I00 funded by FEDER funds provided in the National Spanish projects, the project PID2022-139297OB-I00 funded by MCIN / AEI / 10.13039/501100011033 and by "ERDF A way of making Europe".

REFERENCES

Das, S., Agarwal, N., Venugopal, D., Sheldon, F. T., & Shiva, S. (2020, December). Taxonomy and survey of interpretable machine learning method. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 670-677). IEEE. Retrieved from https://drsaikatdas.com/papers/taxonomy.pdf

David Mermin, N. (1989). What's wrong with this pillow? Physics Today, 42(4), 9-11. Retrieved from https://doi.org/10.1063/1.2810963

Giovanola, B., & Tiribelli, S. (2023). Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. AI & society, 38(2), 549-563. https://doi.org/10.1007/s00146-022-01455-6

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. AI magazine, 40(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850

Marín, G. & Carrasco, R. A. (2021). What Do Machines Think About? In [New] Normal Technology Ethics: Proceedings of the ETHICOMP 2021 (pp. 129-133). Universidad de La Rioja. Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=7977270

Marín, G., Galán, J. J., & Carrasco, R. A. (2022). XAI for Churn Prediction in B2B Models: A Use Case in an Enterprise Software Company. Mathematics, 10(20), 3896. https://doi.org/10.3390/math10203896

Monje, L., Carrasco, R. A., Rosado, C., & Sánchez-Montañés, M. (2022). Deep learning XAI for bus passenger forecasting: A use case in Spain. Mathematics, 10(9), 1428. https://doi.org/10.3390/math10091428

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning. Retrieved from https://scholarspace.manoa.hawaii.edu/bitstreams/d062bd2a-df54-48d4-b27e-76d903b9caaa/download

2. Marketing and Smart Ethics in the digital world

Jesús García de Madariaga, Universidad Complutense de Madrid; Alba García Milon, University of La Rioja; Natalia Medrano Sáez, University of La Rioja; Cristina Olarte Pascual, University of La Rioja; Eva Reinares, Universidad Rey Juan Carlos

EMPOWERING MARKETING ACADEMICS AS INTERDISCIPLINARY KNOWLEDGE INTEGRATORS IN THE FOURTH INDUSTRIAL REVOLUTION

Jesús García-Madariaga, Carlos Lamela-Orcasitas

Economics and Business College, Universidad Complutense de Madrid (Spain)

jesusmadariaga@ccee.ucm.es; clamela@ucm.es

EXTENDED ABSTRACT

On the importance of intelligent data-based systems

Knowledge expressed as the appropriate use of information affects the results of organizations. The creation of knowledge and flexibility in the distribution of information are important for companies, which is why organizations must manage data as the axis on which their strategy for creating economic and social value pivots. This approach is shifting the generation of value from tangible elements to new intangible elements such as skills, information, or knowledge (Vargo & Lusch, 2004).

The application of so-called intelligent systems in marketing is receiving a great deal of attention from academics. Undoubtedly, these topics will continue to be in vogue in the coming years due to the inherent potential to drive change in marketing that information-intensive tools and strategies can bring (Tamaddoni et al., 2014). Considering technological advances, especially the development of artificial intelligence (AI), there are increasing regulatory complexities typical of this fourth industrial revolution. Marketing academics need to focus on improving their skills to become effective knowledge integrators across new frontiers if they are to maintain their role as cutting-edge scientists, needing to understand the ways in which technologies blur the boundaries between spheres of knowledge, so they can participate and even lead interdisciplinary collaborations.

CRM strategies are concerned with creating greater value for the company's shareholders/owners by developing the right relationships with key customers or segments. Given that customers can have negative behaviour towards the company trying to take advantage of it (excessive complaints and requests and through improper use of products and services), it is essential to manage relationships with customers more efficiently, ending with those that are not profitable and choosing better. prospects or potential customers (Oztaysi et al., 2011). Thus, a misunderstanding of the strategic approach by companies can lead to inappropriate exploitation of customers (for example, using intrusive technology), leading to explicit abuse, as CRM technology can provide powerful resources (Frow et al., 2011; Palmer, 2010; Knox, 2003). The appropriate use of information allows specific marketing efforts to be more effective and profitable than massive efforts (Esteban-Bravo et al., 2014). For example, implementing appropriate advertising campaigns according to the precise moment and the appropriate channel, with content tailored to the client's wishes (Kumar et al., 2017). The indiscriminate offer of products and services, without considering preferences, annoys customers and can cause them to lose themselves (Pansari and Kumar, 2017; Tomczyk, 2016). The successful implementation of intelligent CRM systems depends on several factors, such as the intelligent use of data and technologies, the

acquisition and dissemination of customer knowledge among stakeholders, and finally, cooperation within the organization (Bohling et al., 2006).

On the importance of knowing your customers to influence them

Other disciplines of great importance for research on customer behaviour modelling and, therefore, its assessment as key value-generating assets, are behavioural economics and finance, which are already having a significant influence. Individual behaviour depends not only on economic incentives and accessible information, but also on individual preferences, abilities, experiences, and other characteristics. The main conclusion that is derived is that factors at the individual level significantly improve our ability to explain and predict accounting phenomena beyond the company, industry, or sector (Hanlon et al., 2022). Buying behaviour is predominantly influenced by satisfaction and only to a small extent by emotions. The higher the level of positive customer emotions towards the brand, the greater the indirect customer contribution. The higher the customer engagement, the more likely they are to provide the company with access to their personal information and enable them to provide more appropriate and profitable marketing communication (Pansari and Kumar, 2017).

About the importance of regulatory compliance

For all the above, it is necessary to consider the growing relevance that the legal and regulatory dimension that the collection of certain data implies and that may evolve in the future, greatly influencing the way of capturing and managing the necessary information. for the construction of customer valuation models. These issues could become a serious limitation for certain companies or industries and/or countries, especially about specific personal data of customers that may be specially protected by law. This situation means that companies must manage consent for any modelling process that uses personal data and that ethical questions must be raised about how to proceed when collecting information from different sources and how to use it. Working with certain data or using it for certain commercial purposes without the express consent of the clients, apart from the moral dilemmas that it implies for the managers and companies themselves, can lead to substantial sanctions for organizations, both of an economic and reputational nature. Therefore, it is recommended as a good practice that before carrying out work such as the one proposed in this investigation, the pertinent data policy of the organization be reviewed or, where appropriate, developed. Another serious problem facing industries is the significant increase in regulatory and compliance actions by national and international supervisory authorities.

About the importance of business sustainability

Finally, from a holistic perspective of customer data management and its impact on business results, it must be considered that, as society's expectations towards caring for the environment evolve, companies have begun to design strategies to develop sustainable management practices. In the last decade, companies have diverted their focus from purely economic dimensions to also include social and environmental aspects (Madanaguli et al., 2022). For this reason, the analysis of ethical and moral issues that promote the development of sustainable and socially responsible businesses and companies with their environment as demanded by

society today has become essential. Thus, a logical extension of research in customer value modelling is the possible inclusion of metrics, variables or items that consider their value from a sustainability point of view (carbon footprint, energy consumption, etc. This may be especially relevant in certain industries where sustainability plays a determining role (tourism, energy, agriculture, construction, finance, etc.).

KEYWORDS: Ethics, customer valuation, information value, customer relationship management, customer lifetime value.

REFERENCES

Bohling, T., Bowman, D., Lavalle, S., Mittal, V., Narayandas, D., Ramani, G. & Varadarajan, R. (2006). "CRM Implementation: Effectiveness Issues and Insights", *Journal of Service Research*, 9(2),1 84-194.

Esteban-Bravo, M., Vidal-Sanz, J. M. And Yildirim Gökhan (2014). "Valuing customer portfolios with endogenous mass and direct marketing interventions using a stochastic dynamic programming decomposition", *Marketing Science*, 33(5), 621–640.

Frow, P., Payne, A., Wilkinson, I.F. & Young, L. (2011). "Customer management and CRM: addressing the dark side", *Journal of Services Marketing*, Vol. 25, no. 2, pp. 79-89.

Glazer, R. (1991). "Marketing in an information-intensive environment: strategic implications of knowledge as an asset", *Journal of Marketing*, 55(4), 1-19.

Hanlon, M., Yeung, K., & Zuo, L. (2022). "Behavioral economics of accounting: A review of archival research on individual decision makers", *Contemporary Accounting Research*, 39(2), 1150-1214.

Knox, M., Maklan, S., Payne, A., Peppard, J., Ryal, L. (2003). "Customer Relationship Management: Perspective from Market Place", Oxford: Butterworth.

Madanaguli, A., Srivastava, S., Ferraris, A., & Dhir, A. (2022). "Corporate social responsibility and sustainability in the tourism sector: A systematic literature review and future outlook", *Sustainable Development*, 30(3), 447-461.

Oztaysi, B., Sezgin, S. & Ozok, A. F. (2011). "A Measurement Tool for Customer Relationship Management Processes", Industrial Management and Data Systems, 111(6)), 943–960. Palmer, A. (2010). "Customer experience management: a critical review of an emerging idea", *Journal of Services Marketing*, 24(3), 196-208.

Pansari, A. And Kumar, V. (2017). "Customer Engagement: The Construct, Antecedents, and Consequences", *Journal of the Academy of Marketing Science*, 45(3),294–311.

Tamaddoni Jahromi, A., Stakhovych, S. & Ewing, M. (2014). "Managing B2B customer churn, retention and profitability", Industrial Marketing Management, 43(7), 1258-1268. Tomczyk, P. (2016). "Customer knowledge valuation model based on customer lifecycle", *Marketing i Zarządzanie*, 5(46), 87-94.

Vargo. S.L. & Lusch, R.F. (2004). "Evolving to a New Dominant Logic for Marketing", *Journal of Marketing*, 68(1), 1-17.

DARK PATTERNS: TRANSPARENCY OBLIGATIONS AGAINST DECEPTION IN VIRTUAL INFLUENCER MARKETING

Jacopo Ciani Sciolla

Università degli Studi di Torino (Italy)

jacopo.cianisciolla@unito.it

EXTENDED ABSTRACT

'Dark patterns' is an emerging phenomenon in the contemporary attention economy [Davenport (2001), Zuboff (2019)]. The online environment is populated by internet companies exploiting users' psychological vulnerabilities thanks to the use of AI, by coercing, nagging, or deceiving them into making decisions that, if fully informed, they might not make, to maximise profits.

Dark patterns are an umbrella term for manipulative interface design choices that negatively impact the user's decision making, leading the user to act against their interests (e.g., subscribing to a service, purchasing unwanted items, giving away more data than intended). Such practices can amount to consumer [BEUC (2022)] and data protection law violations [EDPB (2022)], e.g., by deceiving users into accepting cookie consent, unwanted purchases or subscriptions, other financial harms, as well as increasing levels of anxiety due to time limits and social pressure.

Given the growing use of dark patterns and the ease with which they can be added to platforms (i.e., dark patterns as a service), the research agenda is strongly focus on the understanding of these practices, consequent harms, and potential countermeasures.

The proposed paper aims to study a phenomenon that the scientific literature does not really address as a dark pattern, but actually has a very similar influence on the end-users from legal and socio-ethical perspectives.

User experience design (UX design) and user interface design are conceptual design disciplines focusing on the interaction between users and machine to design systems and computer interfaces that address the user's experience when using a platform [Dove (2017)]. Good UX revolves around the idea of providing people with interactions that are seamless, enjoyable, and intuitive. To achieve this, a designer should focus on satisfying a user's needs above everything else. However, UX is a tool that could be used for good, or for evil. One such category of evil design is "dark patterns."

The proposed paper shall be focused on another output of UX design, recently taking an even growing stage in the digital environment: the creation of realistic and visually appealing virtual influencer.

Marketing research defines influencers as content creators, who attract the interest of large numbers of consumers on social media platforms. Traditionally, brands collaborate with real-life influencers (i.e., humans living in a physical world) who can make their own decisions regarding sponsored collaborations with brands and form opinions about the products and services they promote.

With recent technological developments, brands increasingly started to work with virtual influencers. Virtual influencers are as non-human digitally created characters sharing social media content and engaging in interactive communications with an aim to obtain influential status among consumers. Within this wide category, experts are used to distinguish between influencers that are created with computer-generated imagery technology (CGI influencers) and AI influencer that rely on artificial intelligence technologies in creating content and interacting with consumers.

Virtual influencers can take different forms and shapes ranging from unimaginable characters that look like simple drawings to hyper-realistic characters that can be nearly impossible to distinguish from real-life (human) influencers. In September, Meta launched 28 AI-powered chatbots featuring avatars of celebrities like Kendall Jenner (Billie), Paris Hilton (Amber), and Snoop Dogg (Dungeon Master). Currently, they are only available for testing in the United States but AI shall make celebrities, in the near future, omnipresent, since they can penetrate every market and formats at any time.

Even though virtual influencers do not exist in 'real' life, several studies showed they are perceived as authentic and 'real' as social media influencers. Consequently, it is not surprising that virtual influencers are capable of being preferred to human.

While virtual influencers have many attractive characteristics for brands, they also raise some concerns. As hyper-realistic virtual humans are designed to have human-like features and behaviours and appear in the physical world, at, for example, real-life restaurants and events, these influencers might be particularly difficult for consumers to distinguish from real-life influencers. Much like deepfakes, the rise of virtual influencers highlights our inability to distinguish reality from fabrications.

Many warn of the serious consequences coming if we can no longer trust any of the information we consume. One day, the prevalence of fake presences may eradicate our sense of reality in the virtual realm.

This risk gets higher and higher when influencers are involved in marketing activities. Here, virtual influencer marketing might suspend consumer's abilities to identify and critically evaluate persuasive marketing tactics. Indeed, there is a danger that owners of virtual influencers withhold information, making consumers falsely believe that they are engaged in communications with humans.

This information is valuable for consumers which pay attention not only to the content that they share but also to who influencers are as individuals. Research shows that consumers are more likely to rely on recommendations from individuals that have views and beliefs similar to their own. Highly anthropomorphic digital characters tend to be perceived as more competent and persuasive as well as to be more successful in developing relationships with consumers.

While there has been some initial research on virtual influencers in law and ethics, such studies are largely descriptive in nature, mostly documenting the existence of these practices.

The main objective of the paper is to address the gap and design a legal and ethical benchmark that would support a clear understanding of the lawful or unlawful nature of virtual influencers marketing and would set the ethical and legal limits to the scope and use of virtual influencer for advertising purposes.

Based on the literature review and evaluation of the applicable hard and soft law, we suggest that brands need to be transparent about using digital characters in their communications through disclaimers. Finally, we advise that when opting to cooperate with virtual influencers, brands should do not engage in marketing communications referring to any testimonial or endorsement of products that would be per se no genuine. The robot's endorsement of the product is in no way based upon its bona fide use, nor is it based upon personal opinions, beliefs, or experiences.

Habermas' theory of communicative action and the Kantian categorical imperative support this opinion.

The proposed paper particularly fits with the "Marketing and Smart Ethics in the digital world" track because virtual influencers are marketing strategies that may be controversial from a legal and ethical standpoint. All in all, the proposed paper will provide a deeper understanding of the phenomenon, often operating in a blurred area between legitimate attempts at persuasion and illegitimate manipulation techniques.

KEYWORDS: Digital marketing, dark patterns, virtual influencer, unfair commercial practices, transparency, ethics.

REFERENCES

BEUC (2022). Dark patterns and the EU consumer law acquis. Recommendations for better enforcement and reform.

Davenport, T.H. & Beck, J.C. (2001). *The Attention Economy: Understanding the New Currency of Business*. Harvard Business School Press.

Dove, G., Halskov, K., Forlizzi, J. & Zimmerman, J. *UX Design Innovation: Challenges for Working with Machine Learning as a Design Material*. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. CHI '17. New York, NY, US: Association for Computing Machinery. pp. 278–288.

EDPB (2022). Guidelines 3/2022 on Dark patterns in social media platform interfaces.

Zuboff, S. (2019) *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Barack Obama's books of 2019.

IS IMPORTANT THE LOSS OF HUMAN CONTACT IN THE ACCEPTANCE OF SOCIAL ROBOTS BY RETAIL CUSTOMERS?

Natalia Medrano, Áurea Subero-Navarro, Jorge Pelegrín-Borondo, Cristina Olarte-Pascual, Eva Reinares-Lara

University of La Rioja (Spain), University of La Rioja (Spain), University of La Rioja (Spain), University of La Rioja (Spain), University of La Rioja (Spain), Universidad Rey Juan Carlos (Spain)

natalia.medrano@unirioja.es, aurea.subero@unirioja.es; jorge.pelegrin@unirioja.es; cristina.olarte@unirioja.es; eva.reinares@urjc.es

EXTENDED ABSTRACT

In recent years, there have been great advances in different disciplines such as computing and mechanics that have enabled the development of robots to perform multiple tasks, both industrial tasks and other tasks such as interation with people in numeous environments (health, education, commercial...) providing services such as care for ederly people, advisory tasks in commercial environments, medical tasks, etc. (Torras, 2014). The increasing digitization of human activity has merged the physical, digital and biological worlds in ways that will change the humanity in its own essence (Porcelli, 2021). In this sense, the different advances are transforming the retail sector (Shankar, 2018; De Bellis & Venkataramani, 2020). Despite the fact that the use of new technologies bring benefits (Grewal et al., 2017), it also raises ehtical dilemmas and it is necessary to formulate modern legislation according to the new reality.

For this reason, there has been an increase in awareness and interest in the ethical considerations for the development of social robots since it is expected that these new technologies will become part of our daily lives in the near future (Malle et al., 2015; Li et al., 2019; Van Maris et al., 2020). All this has been reflected in conferences such as the International Conference on Robot Ethics and Robots Stardards and new ethical Standards in Robotics and AI (Winfield, 2019). Despite this interest on the part of academics, we don't know if it's also interesting for customers that go to stores.

Currently, the main ethical problems that have been most frequently addressed in the literature are privacy/ data control, deception, human autonomy and loss of human contact (Paredo Boada et al., 2021):

- -Privacy/Data control: it's understood as a right against arbitrary interference in one's private life, which leads users to have control over their personal information. Benefits such as reliability and precision are related to this ethical issue.
- -Deception: it's based on the deceptive relationship that human-robot interaction (HRI) can entail. The benefit of reliability is related to this ethical problem.
- -Autonomy: excessive use of technology could lead to a loss of users capabilities. The benefit of not relating to human beings is related to this ethical problem.
- -Loss of human contact: the use of social robots could enhace social isolation. The benefit of not relating to human beings is algo related to this ethical issue.

Research questions, objectives and methodology

Social Robots research in retail represents a new field of marketing research given it's disruptive and distinctive characteristics. There is an important gap in the literature (Belanché et al., 2020; Grewal et al., 2017) that needs to be answered: Are ethical aspects important for customers to decide to use social robots?

That's why our research addresses the potential core benefits sought from using social robots in trading through a sequential proccess. First, a qualitative analysis was carried out based on an open question asked to 1.069 individuals over 18 years of age that live where the 12 main basic benefits sought were obtained (see table 1). Several of these main basic benefits in the acceptance of social robots are related to ethical aspects related to human contact: i) if a robot serves me, i will avoid possible unpleasant treatment by sellers; iii) if a robot assists me, i will avoid interacting with the sellers; iv) if a robot serves me, i will have the same treatment as the rest of the customers (i will avoid discrimination).

After this, secondly, with a second sample of 735 individuals over 18 years of age residing in Spain, the model was contrasted by applying a personal survey on the benefits that have been obtained in the qualitative analysis and the intention to use social robots.

Table 1. Expected Benefits.

	Average size	**Variance**
Expected Benefits	If a robot assists me in the store,I will have more reliable information	0 1 2 3 4 5 6 7 8 9 10
	If a robot assists me in the store, my purchase will be more confortable	0 1 2 3 4 5 6 7 8 9 10
	If a robot assists me in the store, my purchase will be faster	0 1 2 3 4 5 6 7 8 9 10
	If a robot assists me in the store, my purchase will be easier	0 1 2 3 4 5 6 7 8 9 10
	If a robot assists me in the store, my purchase will be pleasant	0 1 2 3 4 5 6 7 8 9 10
	If a robot assists me in the store, my purchase will be more accurate (without errors)	0 1 2 3 4 5 6 7 8 9 10
	If a robot assists me in the store, I will be able to buy at any time	0 1 2 3 4 5 6 7 8 9 10
	If a robot assits me in the store, they will be able to low the prices and I will buy cheaper	0 1 2 3 4 5 6 7 8 9 10
	If a robot assists me in the store, I will solve my accessibility problems (example: language, mobility, hearing…)	0 1 2 3 4 5 6 7 8 9 10
	If a robot assists me in the store, I will avoid possible unpleasant treatment	0 1 2 3 4 5 6 7 8 9 10
	If a robot assists me in the store, I will avoid to interact with sellers	0 1 2 3 4 5 6 7 8 9 10
	If a robot assists me in the store, I will have the same treatment as the rest of the customers	0 1 2 3 4 5 6 7 8 9 10

Results

The results of the exploratory factor analysis showed an adequate KMO (0.933) and the efficiency of Barlett's test reflected a significance level <0.001. Based on the results, a confirmatory factor analysis was performed. We obtained 3 factors. The goodness of fit results for the confirmatory were satisfactory: BBNFI=0.958; BBNNFI=0.953; CFI=0.965; robust CFI=0.971; GFI=0.947; AGFI=0.915; RMSEA=0.078; robust RMSEA=0.067. All the variables showed loads higher than 0.7 except two of them that showed values slightly lower than 0.7. However, all had t-values>0.96.

In terms of reliability and convergent validity, it's adequate. Regarding the convergent validity criteria, the average variance extracted (AVE) of all the constructs was greater than 0.5 for the factors (Hair, Anderson, Tatham y Javis, 2005).

Table 2. Composite Reliability and AVE.

Factor	Composite Reliability	AVE
1	0,923	0,668
2	0,785	0,549
3	0,720	0,563

In terms of discriminant validity, all possible correlations between the factors have been calculated. In this way, the confidence interval of the correlations between the dimensions has been obtained. As it's shown in Table 4, the discriminant validity can be supported since none of the confidence intervals of these correlations contains the value 1. Therefore, there is no covariance problem between the factors involved and the discriminant validity test is achieved.

Table 3. Discriminant validity. Confident intervals of the correlations between the dimensions.

	Covariance	Standard Error	Int. conf. Covar.		Interval conf. Correlat.		Var 1 er Fac	Var 2º Fac
F2-F1	0,695	0,026	0,643	0,747	0,643	0,747	1	1
F3-F1	0,782	0,025	0,732	0,832	0,732	0,832	1	1
F3-F2	0,923	0,023	0,877	0,969	0,877	0,969	1	1

After this verification of the dimensions in which the basic benefits were grouped, factor 3 included the dimension related to the ethical aspects of human contact. Subsequently, using Structural Equations based on Covariances, the influence of the three factors on the intention to use social robots was analysed. The model fits were good: BBNFI = 0.96; BBNFI Robust= 0.96; BBNNFI= 0.95; CFI = 0.96; robust CFI = 0.97; AGFI = 0.90; RMSEA = 0.076; ; robust RMSEA=0.067. The model showed an R2 = 0.65. The first factor showed a path coefficient = 0.66 and significant, the second factor a path coefficient = -0.103 and not significant and the third factor a path coefficient = 0.270 and not significant.

The factor related to ethical aspects (second factor) did not show a significant influence on the intention to use social robots. The factor that influenced the intention to use social robots (factor 1) was related to the fact that the use of social robots makes shopping more reliable, comfortable, easy, pleasant, fast and precise. Therefore, related to usefulness of using the social robot.

Conclusions

The current focus of the retail industry is mainly based on the transformation of the point of sale through the use of technology (Paschen et al., 2019) and the autonomy of customer service (Baird, 2018). Social Robots are going to transform the current shopping experience. Taking into account the benefits obtained from the qualitative analysis, the ethics related to social contact can be a relevant point in the acceptance by customers of social robots in retail commerce. Nowadays there is a notable lack of attention regarding the implications of robotics from a practical point of view. That is, how is the practice being tranformed when a robot is introduced? (Pareto Boada, 2021). If robotics is introduced into everyday tasks, it's necessary to think on the implications according to the values and purposes of the practice it serves. Our results show that despite the importance given by the literature to ethics on aspects of human contact, we have seen that the factor related to ethical aspects associated to human contact, even though it's a basic benefit by customers, it's not a determining factor when using social robots in retail.

KEYWORDS: Social robots, smart technologies, ethics, technology acceptance.

ACKNOWLEDGEMENTS: The authors gratefully acknowledge the support from the University of La Rioja for funding given to the COBEMADE Research Group at the University of La Rioja.

REFERENCES

Baird, N. (2018, June 10). Robots, automation and retail: not so cut and dried. *Forbes*. Retrieved from: https://www.forbes.com/sites/nikkibaird/2018/06/19/robots-automation-and-retail-not-so-cut-and-dried/?sh=295210eb7b06.

Belanché, D., Casaló, L.V., Flavián, C., & Schepers, J. (2019). Service robot implementation: a theoretical framework and research agenda. *The service Industries Journal,* 40(3-4), 203-225.

De Bellis, E., & Venkataramani, G. (2020). Autonomous Shopping Systems: Identifying and overcoming barriers to consumer adoption. *Journal of Retailing,* 96(1), 74-87.

Grewal, D., Roggeveen, A., & Nordált, J. (2017). The future of retailing. *Journal of Retailing,* 93(1), 1-6.

Hair, H.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (2005). *Análisis Multivariante*. Ed. Prentice Hall, Madrid.

Li, H., Milani, S., Krishnamoorthy, V., Lewis, M., & Sycara, K. (2019). Perceptions of domestic robots normative behavior across cultures. *Proceedings of the 2019 AAAI/ACM Conference on Artificial Intelligence, Ethics and Society,* 345-351.

Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. *Proceedings of 2015 10th ACM/IEEE International Conference on Human Robot Interaction*, 117-124.

Paredo Boada, J. (2021). Prolegómenos a una ética para la robótica social. *Dilemata*, 34, 71-78.

Paschen, J., Kietzmann, J., & Kietzmann, T.C. (2019). Artificial intelligence (AI) and its implications for market knowledge in B2B marketing. *Journal of Business & Industrial Marketing*, 1-10.

Porcelli, A.M. (2020). La inteligencia artificial y la robótica: sus dilemas sociales, éticos y jurídicos. *Derecho global, Estudios sobre Derecho y Justicia,* 16, 49-105.

Shankar, V. (2018). How artificial intelligence (AI) is reshaping retailing. *Journal of Retailing*, 94(4), vi-xi.

Torras, C. (2014). Robots Sociales. Un punto de encuentro entre ciencia y ficción. *MÉTODE Science Studies Journal*, 4, 1-5.

Van Maris, A., Zook, N., Caleb-Solly, P., Winfield, A., & Dogramadzi, S. (2020). Designing Ethical Social Robots-A longitudinal Field Study with older adults. *Frontiers in Robotics and AI*, 1-14.

Winfield, K., & Van Maris, A. (2019). Social influence and deception in socially assistive robotics. *Proceedings of the International Conference on Robots Ethics and Standards (ICRES 2018),* 1-2.

AR AND VR IN THE SPOTLIGHT: A SYSTEMATIC LITERATURE REVIEW OF SECURITY, PRIVACY, AND ETHICAL CONCERNS

Alba García-Milon, Mandy tom Dieck

University of La Rioja (Spain), Manchester Metropolitan University

alba.garciam@unirioja.es; C.tom-Dieck@mmu.ac.uk

EXTENDED ABSTRACT

Virtual Reality (VR) and Augmented Reality (AR) are two immersive technologies that have gained a great interest in the business field and, as a result, are being implemented in many business-based activities (Cranmer et al., 2020; Han et al., 2019). AR allows to explore unknown surroundings in an interactive, informative and enjoyable way (Cranmer et al., 2020) and VR facilitates virtual visitation to environments from anywhere and anytime and becomes a resource able to transmit the experience and intangibility of spaces to the user (Huang et al., 2016).

Despite the great benefits and potential that AR and VR are showing for the industry, serious security and privacy concerns have been identified (Guzman et al., 2019; Lebeck et al., 2018). For instance, the possibility of recording sensitive information from the user's surroundings or interfere in the user's view of the environment are some of the main risks associated with AR according to Lebeck et al., (2018). However, up to the present time, there is no SLR that emphasizes a comprehensive understanding of privacy, security and ethical concerns in extended reality, an issue of foremost importance in the adoption of this technology. Consequently, the current study starts covering this gap by summarizing, analyzing, and synthesizing the relevant corpus of literature that arises issues of privacy, security and ethical concerns in extended reality focusing on business and management field. This is an issue that raises new research questions and still needs to be addressed by academia (Ameen et al., 2021). The purpose of the current study is to explore the AR/VR privacy and security concerns in business employing a systematic literature review (SLR) method.

SLR is pertinent for the identification, selection, analysis and evaluation of the relevant literature (Mohamed Shaffril et al., 2021; Tranfield et al., 2003). Its objective is to analyse the relations, contradictions, and gaps among the results of all the shortlisted literature and it represents an optimal start point to provide suggestions for future research (Jain et al., 2022). In order to conduct the present SLR, a structured process following the guidelines provided by Denyer and Tranfield (2009) was conducted. This rigorous process enables the search for all studies that are potentially significant (Denyer & Tranfield, 2009). The mentioned process for SLR has been recently employed in the business field (e.g. Heinis et al., 2021; Jain et al., 2022) with satisfactory results. This process includes five refinement stages: Question Formulation, Locating Studies, Study Selection and Evaluation, Analysis and Synthesis and Reporting and Using the Results.

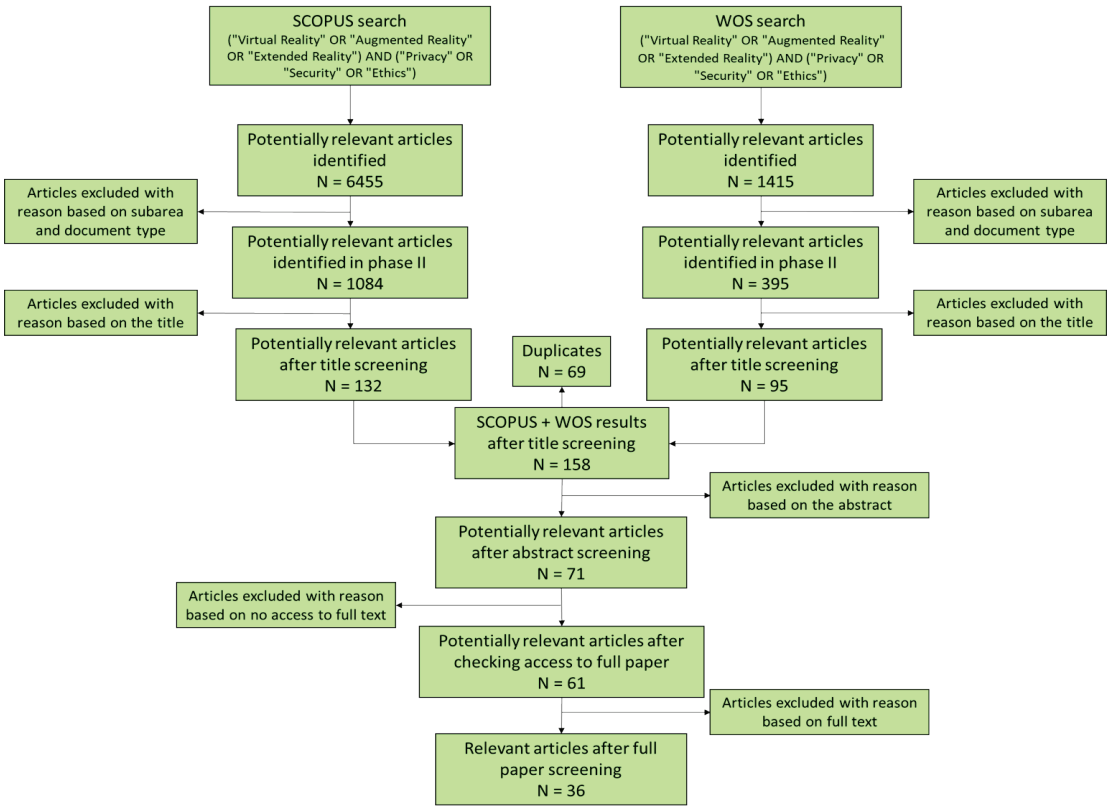
Regarding the first stage, question formulation, the selected question to be answered was: "What are the privacy, security and ethical concerns to adopt extended reality in the business field?". Following with the second stage, locating studies, 2 search engines were employed (Scopus and WOS) and the terms used for the search were "Virtual Reality" OR "Augmented

Reality" OR "Extended Reality") AND ("Privacy" OR "Security" OR "Ethics". This stage was made in December of 2021. Figure 1 shows the screening process followed for locating and selecting studies to answer the formulated question:

In the initial analysis of the selected articles, some notable trends and patterns have emerged. Firstly, a substantial portion of the articles were published in 2021, indicating a recent surge in research interest in the field of privacy, security, and ethical concerns within the AR and VR context. This suggests that the topic is gaining increasing attention and relevance within the academic community.

Of the articles reviewed, approximately 53% were empirical studies, while the remaining 47% were theoretical in nature. This distribution reflects a balanced mix of research approaches, enabling a comprehensive examination of the subject matter. The inclusion of empirical studies indicates a focus on gathering real-world data and insights related to security, privacy, and ethics in AR and VR applications, offering valuable insights into practical implications and experiences.

Figure 1. Screening process.



Regarding the context of the selected articles, it was observed that the most common context explored was technology, with 12 papers. This emphasis on technology highlights the significant impact that AR and VR has on various industries and the need to address associated risks and considerations. Following technology, the second most prevalent context was retail, with 7 papers. This finding suggests that AR and VR technologies hold promising potential for transforming the retail industry, prompting researchers to critically analyze the implications of implementing AR and VR in this domain.

Additionally, a smaller number of papers, 3 in total, focused on marketing. This indicates that researchers are beginning to explore how AR and VR can shape marketing practices and the potential challenges associated with safeguarding consumer privacy and maintaining ethical standards in AR/VR-driven marketing campaigns.

Moving forward, the next steps will involve a deeper exploration of the content within the identified studies. This includes critically examining and problematizing the literature to gain a comprehensive understanding of the security, privacy, and ethical concerns within the AR and VR context. By analyzing and synthesizing the existing research, the study aims to identify gaps, inconsistencies, and unresolved issues within the current body of knowledge. Moreover, the study seeks to identify future research directions and areas of exploration for AR and VR in the business and management field. This entails identifying key challenges, emerging trends, and potential opportunities that warrant further investigation. By doing so, the research aims to contribute to the advancement of knowledge in the field, providing valuable insights for practitioners, policymakers, and scholars interested in the intersection of AR and VR, privacy, security, and ethical considerations in the business and management domain.

KEYWORDS: Virtual reality, augmented reality, security, privacy, ethical concerns.

REFERENCES

Ameen, N., Hosany, S., & Tarhini, A. (2021). Consumer interaction with cutting-edge technologies: Implications for future research. Computers in Human Behavior, 120(August 2020), 106761. https://doi.org/10.1016/j.chb.2021.106761

Cranmer, E. E., tom Dieck, M. C., & Fountoulaki, P. (2020). Exploring the value of augmented reality for tourism. Tourism Management Perspectives, 35(March). https://doi.org/10.1016/j.tmp.2020.100672

Denyer, D., & Tranfield, D. (2009). Producing a systematic review. In D. A. Buchanan & A. Bryman (Eds.), *The Sage Handbook of Organisational Research Methods* (pp. 671–689). London: Sage.

Guzman, J. A. D. E., Thilakarathna, K., & Seneviratne, A. (2019). Security and privacy approaches in mixed reality: A literature survey. ACM Computing Surveys, 52(6). https://doi.org/10.1145/3359626

Han, D. I. D., Jung, T., & Tom Dieck, M. C. (2019). Translating Tourist Requirements into Mobile AR Application Engineering Through QFD. International Journal of Human-Computer Interaction, 35(19), 1842–1858. https://doi.org/10.1080/10447318.2019.1574099

Heinis, S., Bamford, D., Papalexi, M., & Vafadarnikjoo, A. (2021). Services procurement: A systematic literature review of practices and challenges. *International Journal of Management Reviews*, *July 2019*, 1–21. https://doi.org/10.1111/ijmr.12281

Huang, T. L., & Liao, S. (2015). A model of acceptance of augmented-reality interactive technology: the moderating role of cognitive innovativeness. Electronic Commerce Research, 15(2), 269–295. https://doi.org/10.1007/s10660-014-9163-2

Jain, R., Jain, K., Behl, A., Pereira, V., Del Giudice, M., & Vrontis, D. (2022). Mainstreaming fashion rental consumption: A systematic and thematic review of literature. *Journal of Business Research*, *139* (March 2021), 1525–1539. https://doi.org/10.1016/j.jbusres.2021.10.071

Lebeck, K., Ruth, K., Kohno, T., & Roesner, F. (2018). Towards Security and Privacy for Multi-user Augmented Reality: Foundations with End Users. Proceedings - IEEE Symposium on Security and Privacy, 2018-May, 392–408. https://doi.org/10.1109/SP.2018.00051

Mohamed Shaffril, H. A., Samsuddin, S. F., & Abu Samah, A. (2021). The ABC of systematic literature review: the basic methodological guidance for beginners. *Quality and Quantity*, *55*(4), 1319–1346. https://doi.org/10.1007/s11135-020-01059-6

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management*, *14*(3), 207–222. https://doi.org/10.1111/1467-8551.00375

3. Open Track

Kiyoshi Murata, Meiji University; Ana María Lara Palma, Universidad de Burgos; Yohko Orito, Ehime University

EU AI ACT AND ITS CONDITIONS FOR HUMAN FLOURISHING: A VIRTUE ETHICS PERSPECTIVE

Salla Westerstrand

Turku School of Economics, University of Turku (Finland)

sakrpon@utu.fi

EXTENDED ABSTRACT

Recent developments in Artificial Intelligence (AI) have accelerated initiatives to guide and regulate the use thereof. In the European Union (EU), the European Commission (Commission) gave its proposal for a regulation of the European Parliament and of the proposal for a Regulation of European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) (hereinafter: EU AI Act) in April 2021 (European Commission 2021). The proposal was followed by intense political discussions and amendments, leading to the adoption of the General Approach of the Council of the EU in November 2022 and amendments adopted by the European Parliament in June 2023. Currently, the draft is in trilogue negotiations between the three EU institutions and is scheduled to be adopted by the end of 2023.

Whereas the upcoming EU AI Act is not an ethics guideline[2], it is of high relevance when evaluating the societal conditions that guide the ethical directions of AI systems development in Europe but also globally. First, unlike the Ethics Guidelines for Trustworthy AI published by the Commission in 2019 (HLEG, 2019), the EU AI Act is legally binding and introduces sanctions for providers and users of the regulated AI technologies upon non-compliance. Therefore, it can be expected that the EU AI Act contributes to setting the basis for ethical directions of AI development.

Second, it addresses certain phenomena that have been shown to be subject to ethical dilemmas. We could ask, for instance: is it justified to exercise effective surveillance in public spaces and hence limit human freedom to protect citizens' physical safety against a terrorist attack (Almeida et al., 2022; Zuboff, 2019)? Can it be acceptable to deploy highly accurate algorithms in recruitment and thus find better fitting workplaces for most people more efficiently, if that means a higher risk for systematic discrimination of minorities (for a review of ethics of AI in recruitment, see Hunkenschroer & Luetge, 2022)? Both are situations that fall into the scope of the EU AI Act. It includes prohibitions to AI-powered surveillance and deems AI used in recruiting high-risk and thus subject to further requirements. Furthermore, if the EU AI Act leads to the adoption of corresponding measures in non-EU countries similar to the case of the General Data Protection Regulation (GDPR) (a.k.a. *Brussels effect*), it is anticipated to have a global impact on the direction which AI is being developed (Siegmann & Anderljung, 2022). Hence, despite arguably being a Eurocentric perspective and thus not reflecting the full global discourse around AI development, it is expected to offer a fruitful starting point for better

[2] The Commission gave its recommendations, Ethics Guidelines for Trustworthy AI, in 2019, which is available here: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

understanding the ethical directions of ongoing AI development, mainly in the EU but also beyond.

Still, is compliance with the EU AI Act a guarantee or even a promise of ethical AI? As the latest versions of the EU AI Act were introduced only recently, we still lack understanding of what kind of ethical implications the European regulation could have on AI development when adopted. Understanding the ethical directions implied by the EU AI Act is essential for several reasons: First, it offers the providers of AI systems, namely IS practitioners, a better understanding of what can be achieved by EU AI Act compliance in terms of ethical AI, and what perhaps falls out of the scope of the European legislation and needs to be done by other means. Second, it offers guidance for policymakers to fill the gaps left behind by the EU AI Act in order to work towards more ethical AI systems. Lastly, it offers ground for IS researchers in academia and the private sector to start exploring practical methods and solutions for mitigating ethical issues in AI systems.

Hence, to work towards this deeper understanding and ethical AI systems, we seek a response to the following research question:

What kinds of ethical directions does the EU AI Act imply for AI development?

In this paper, we approach the question from the perspective of virtue ethics, which is a much-discussed branch of moral philosophy originating from the work of Aristotle. We concentrate on the perspective introduced by Bynum (2006), inspired by Norbert Wiener, James Moor, and Luciano Floridi. According to Bynum, only human flourishing can create conditions for ethical action. This perspective of virtue ethics has been shown to be relevant in the context of AI ethics (Stahl, 2021; Stahl et al., 2022; Stenseke, 2021), which makes it a theoretically interesting starting point for analysing the upcoming EU legislation. As Stahl et al. (2022) demonstrate in their study that analyses the role of a European Agency for AI introduced in the EU AI Act, flourishing ethics can serve as a fruitful perspective to shed light on the ethical implications of the upcoming European legislation.

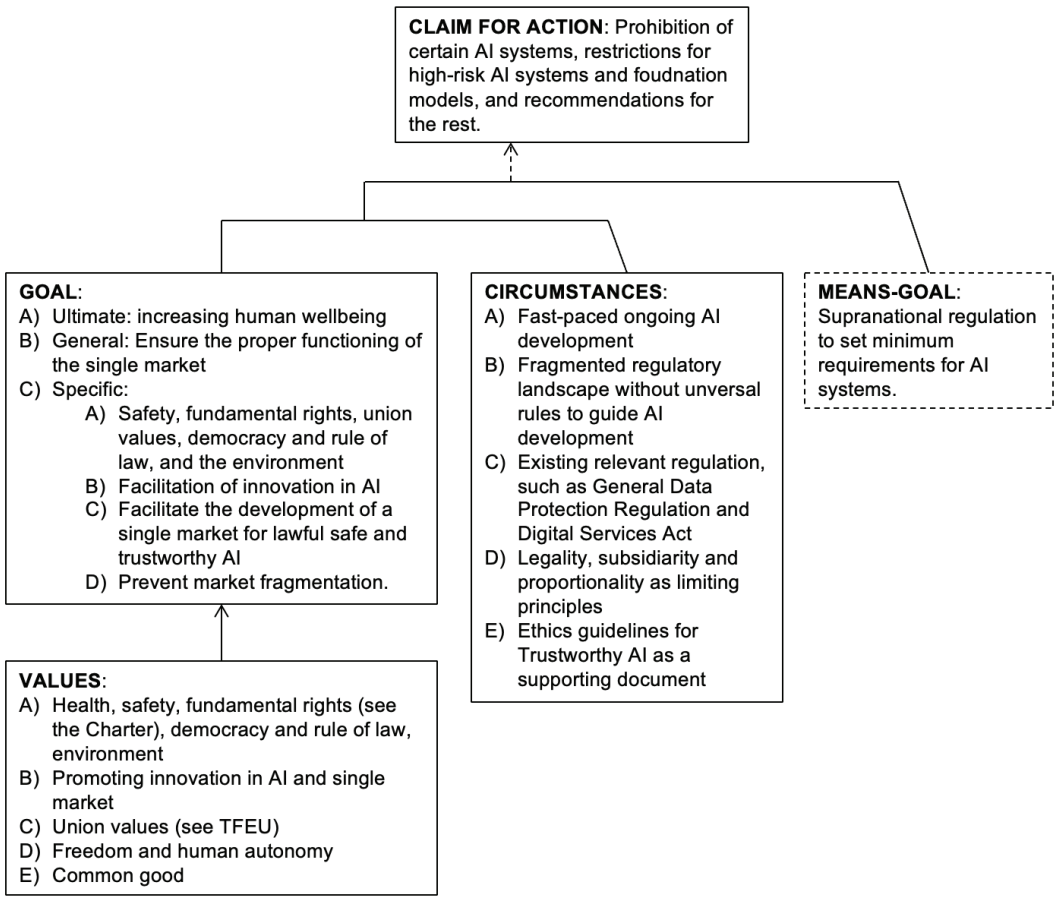
The analysis is conducted using the three latest and major versions of the EU AI Act, i.e., the original proposition by the European Commission, the General Approach adopted by the Council of the European Union, and the amendments adopted by the European Parliament in June 2023, as they are assumed to give the most complete picture available of the main elements of the EU AI Act at the time of writing. The analysis is thus conducted before the results of the trilogue negotiations between the Parliament, Commission, and the EU Council are finalised.

The Act is analysed from the perspective of critical theory, using methodologies of Critical Discourse Studies and Political Discourse Analysis. Critical theories share concern around freedom, autonomy, and human emancipation (Adorno & Horkheimer, 1979), which is often reflected in studies as an interest towards power relations in society (Van Dijk, 2017; Waelen, 2022). They highlight the pragmatic nature of science and knowledge, aiming not only to describe but also to change society by challenging existing paradigms and suggesting alternative courses of action (Delanty & Harris, 2021; Orlikowski & Baroudi, 1991; Stahl, 2008; Waelen, 2022). Approaching ethics in the EU AI Act from this perspective paves the way for justified critique and suggestions for practitioners, policymakers, and deployers of AI technologies, which contributes to directing IS development in a more ethical direction. Such an aim can be considered desirable if not a responsibility for an IS researcher (Chiasson et al., 2018).

Furthermore, we believe that in the age of increasing complexity noted by (Benbya et al., 2020), and the fast-paced development of new systems and user interfaces, offering such guidance helps with ensuring that both AI practitioners and policymakers are informed by scientific research when defining the conditions and technical specifications that shape people's lives.

We build upon the principles for critical IS research offered by Myers & Klein (2011), and critical discourse studies based on Jürgen Habermas's work. We structure the ethics discourse in the EU AI Act around the argumentation scheme introduced in Political Discourse Analysis by Fairclough and Fairclough (2013). The argumentation of the EU AI Act can be illustrated in Figure 1.

Figure 1. Argumentation of the EU AI Act, following the structure of practical arguments by Fairclough and Fairclough (2013).



According to Bynum (2006), the central elements for human-centred flourishing ethics derived from Aristotle's virtue ethics are the following:

- Human flourishing is at the centre of ethics.
- Humans are social, and hence human flourishing requires a connection to society.
- In order to flourish, humans need to do what they are best equipped to do.

- Flourishing requires humans to reason theoretically and practically using intellect and practical judgment. This leads to acting according to the reasoning, in a societal context.
- The key to mastering practical reasoning, and thus to being ethical is ”the capacity to deliberate well about one’s overall goals and carry out that action.”

Analysing the EU AI Act through the lens of these elements demonstrates that the EU AI Act is likely to have impacts on a) human autonomy through practical reasoning, intellect, and deliberation about one’s goals; b) connecting with society; and c) doing what humans are best equipped to do, all of which contribute to human flourishing. The present study shows an acute need for further research concerning the ethical implications of AI regulation from several different ethical perspectives. We hope that the results of the present paper encourage researchers, policymakers, and industry professionals to explore the implications of their AI systems that fall outside the mandate and political emphases of the EU legislation in order to strive towards more ethical future AI systems.

KEYWORDS: Artificial Intelligence ethics, virtue ethics, human flourishing, EU AI Act, ethics of AI regulation.

REFERENCES

Adorno, T. W., & Horkheimer, M. (1979). *Dialectic of Enlightenment*. Verso.

Almeida, D., Shmarko, K., & Lomas, E. (2022). The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: A comparative analysis of US, EU, and UK regulatory frameworks. *AI and Ethics*, *2*(3), 377–387. https://doi.org/10.1007/s43681-021-00077-w

Benbya, H., Nan, N., Tanriverdi, H., & Yoo, Y. (2020). *Complexity and Information Systems Research in the Emerging Digital World* (SSRN Scholarly Paper No. 3539079). https://papers.ssrn.com/abstract=3539079

Bynum, T. W. (2006). Flourishing Ethics. *Ethics and Information Technology*, *8*(4), 157–173. https://doi.org/10.1007/s10676-006-9107-1

Chiasson, M., Davidson, E., & Winter, J. (2018). Philosophical foundations for informing the future(S) through IS research. *European Journal of Information Systems*, *27*(3), 367–379. https://doi.org/10.1080/0960085X.2018.1435232

Delanty, G., & Harris, N. (2021). Critical theory and the question of technology: The Frankfurt School revisited. *Thesis Eleven*, *166*(1), 88–108. https://doi.org/10.1177/07255136211002055

European Commission (2021). Proposal for Regulation of the European Parliament and of the Council.

Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM/2021/206 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

Fairclough, I., & Fairclough, N. (2013). *Political Discourse Analysis: A Method for Advanced Students*. Routledge.

HLEG. (2019). *Ethics Guidelines for Trustworthy AI*. High-Level Expert Group on AI by the European Commission.

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

Myers, M. D., & Klein, H. K. (2011). A Set of Principles for Conducting Critical Research in Information Systems. *MIS Quarterly*, *35*(1), 17–36. https://doi.org/10.2307/23043487

Orlikowski, W. J., & Baroudi, J. J. (1991). Studying Information Technology in Organizations: Research Approaches and Assumptions. *Information Systems Research*, *2*(1), 1–28. https://doi.org/10.1287/isre.2.1.1

Siegmann, C., & Anderljung, M. (2022). *The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market* (arXiv:2208.12645). https://doi.org/10.48550/arXiv.2208.12645

Stahl, B. C. (2008). The ethical nature of critical research in information systems. *Information Systems Journal*, *18*(2), 137–163. https://doi.org/10.1111/j.1365-2575.2007.00283.x

Stahl, B. C. (2021). Concepts of Ethics and Their Application to AI. In B. C. Stahl (Ed.), *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies* (pp. 19–33). Springer International Publishing. https://doi.org/10.1007/978-3-030-69978-9_3

Stahl, B. C., Rodrigues, R., Santiago, N., & Macnish, K. (2022). A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values. *Computer Law & Security Review*, *45*, 105661. https://doi.org/10.1016/j.clsr.2022.105661

Stenseke, J. (2021). Artificial virtuous agents: From theory to machine implementation. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01325-7

Van Dijk, T. A. (2017). *Discourse and Power*. Bloomsbury. https://www.bloomsbury.com/us/discourse-and-power-9780230574090/

Waelen, R. (2022). Why AI Ethics Is a Critical Theory. *Philosophy & Technology*, *35*(1). https://doi.org/10.1007/s13347-022-00507-5

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. https://www.hachettebookgroup.com/titles/shoshana-zuboff/the-age-of-surveillance-capitalism/9781610395694/?lens=publicaffairs

THE COUNTERVAILING POWER OF AI DAOS INFLUENCES VALUE TRANSFORMATION; BITCOIN (POW) VS. ETHEREUM (POS)

Kazuyuki Shimizu

School of Business Administration, Meiji University (Japan)

shimizuk@meiji.ac.jp

EXTENDED ABSTRACT

In this paper, we investigate the process of value transformation influenced by the countervailing power of decentralised autonomous organisations (DAO) controlled by artificial intelligence (AI). There are various values in society. The Internet was believed to lead humanity better by further decentralizing various values. However, the uneven distribution of information causes many problems, such as "cyber cascades", "filter bubbles", and "echo chambers" etc. To solve these problems, DAO using blockchain is expected to become a method to solve those problems in the Internet space.

This paper intends to capture the value change in three steps. In the first step, we take two philosophical approaches. First, using Hegel's dialectic, we attempt to compare A. current social values (thesis), B. the value of, for example, AI DAOs (antithesis), and C. new values (synthesis) in an Internet world where these contradictions exist (A + B = C). The second philosophical approach is to consider the mutuality of Bitcoin (POW) and Ethereum (POS) by adapting Popper's World 1, 2, and 3 models.

In the second step, we examine the three core capabilities that amplify "credit" in decentralised finance (DeFi): exchange swap rates, staking, and indices (portfolio). Finally, we consider the value dispersion within the Bitcoin and Ethereum networks as the main two poles. The countervailing power between these two poles and the upcoming expanding AI DAOs, coordinate the interests of stakeholders.

I. METHODOLOGY

Humanity has various values: legitimacy, human rights, freedom, security, dignity and human life. First, consider two methodologies. Hegel's dialectics; For example, consider value relationships using this dialectic. Also, we applied Popper's World 1/2/3 theory to the current value relationships.

- Based on **Hegel's dialectics**, A. Blockchain "openness" (thesis) + B. "privacy" (antithesis) = C. "improvement of decision" (synthesis).
- **Popper's Falsifiability and Worlds I, II. III**; the "openness" of blockchain technology (world 1), the dilemma between World 1 and "privacy" (world 2), and the new Web 3 interaction with now captured values (world 3). (Karl, 1978)

The global crypto market cap is $1.16T on June 2023. As mentioned above, Bitcoin's dominance is around half of the total crypto market (46.53% in June 2023). The share of Ethereum is around 19% (Omkar, 2023). The total crypto market trading volume is $32.26B. The total volume in DeFi

is about 8% ($2.37B), and the volume of all stablecoins is about 92% ($29.69B) of the total crypto market. We will examine the concept here with POW, POS and the rest of the consensus algorithms for simplicity and understanding.

Diagram of knowledge growth

P_1 *(Problem 1) – TT (Tentative Theory) – EE (Error Elimination) – P (Problem 2)*

- Based on Hegel's dialectics, A. Bitcoin's POW (thesis) + B. Ethereum's POS (antithesis) = C. co-evolution of both consensus algorithms (synthesis). The simple formula is A + B = C.
- Popper's Falsifiability and Worlds I, II. III; Bitcoin's POW (world 1), Ethereum's POS (world 2), and co-evolution of both consensus algorithms (world 3). Bitcoin's POW (world 1), Ethereum's POS (world 2), and co-evolution of both consensus algorithms, such as NPoS (Nominated Proof of Stake), PoH (Proof of Histrory), PoA (Proof of Authority), PoC (Proof of Consensus) etc (world 3).

II. DEFI (DECENTRALISED FINANCE) AS A CREDIT AMPLIFICATION FUNCTION

II.1. Brief history of Defi

What we look for in a financial institution is trust. For this reason, financial institutions are also called credit institutions. Financial institutions are trusted because they can settle transactions smoothly. The source of trust in financial institutions is that they do not lie. A similar trust is formed if this source of trust is unbreakable Cryptography. The POW blockchain project, Bitcoin, secures this trust.

Personal overseas money transfers and on-demand transactions became possible, enabling several online value transformations. As a result, cryptocurrencies have significantly impacted traditional centralised financial (Cefi). However, decentralised finance (Defi) in the early 2000s was still in the prototype technologically, and its business application had just begun (Nakamoto, 2008). Ethereum was developed by Vitalik Buterin in 2013 as the research and development of blockchain technology progressed (Buterin, 2014). By incorporating smart contracts based on cryptocurrencies, blockchain technology can be used for various purposes.

Furthermore, by changing Bitcoin's Proof of Work (PoW) problem to Proof of Stake (PoS), the problem of energy consumption can now be addressed. By decentralising such diverse values, the Internet was believed to lead society in a more pluralistic and better direction. However, information is unevenly distributed, causing problems such as "cyber cascades," "filter bubbles," and "echo chambers mentioned above." To solve such issues, DAO on Web3, powered by blockchain, is expected to promote the democratisation of the "human" internet space (Simon, 2023).

II.2. Orderbook vs Liquidity Pool at Leyer 1

Uniswap is a decentralised exchange (DEX) and part of the decentralised finance (Defi) product ecosystem launched on Ethereum mainnet in November 2018. It replaces the traditional order book type trading on centralised exchanges (CEX). At a very high level, an AMM (Automated Market Maker) replaces the buy and sell orders in an order book market with a liquidity pool of two assets valued relative to each other. As one asset is traded for the other, the relative prices

of the two assets shift and a new market rate for both is determined. In this dynamic, a buyer or seller trades directly with the pool rather than with specific orders left by other parties (Uniswap, 2023). Liquidity providers can be regarded as investors in the decentralised exchange and earn fixed commissions per trade. They lock up funds in liquidity pools for distinct pairs of currencies allowing market participants to swap them using the improved price function**.** Liquidity providers take on market risk as liquidity providers in exchange for earning commissions on each trade. In short, Investors as a customer accept market risks usually taken by traders. This new pool trading has a risk profile of a liquidity provider and the so-called impermanent (unrealised) loss in particular (Andreas & Gurvinder, 2021).

Table 1 below explains the following; **Uniswap** introduced the constant product market maker formula to ensure continuous liquidity in exchanging tokens on Ethereum. The formula follows: x is token 1, y is token 2, and k is a constant.

Curve's primary distinction from other decentralised exchanges, such as Uniswap, is its low slippage and low fee algorithm specifically designed for trading between assets of the same value. Curve makes it very useful for stablecoin swaps, as they are expected to hold roughly the same value [Perpetual Protocol, 2022].

Balancer's pools are like index funds that construct a portfolio of assets with fixed weights. The balancer protocol allows each pool to have two or more assets and to supply them in any ratio. Each asset reserve is given a weight when the pool is created and the weights sum to one.

Table 1. A pattern of AMM.

AMM	**Connection of individual tokens**
Uniswap	X * Y = K
Balancer	(X * Y * Z) ^ (1/3) = K
Curve	(X * Y) + (X + Y) = K

Source: JBA Defi study group.

III. AI DAO AND ARTIFICIAL SOCIETIES

Following Hegel's dialectics, we consider the current values (thesis), the AI DAOs values (antithesis), and the new values (synthesis) in which the two are in a countervailing relationship in the Internet world (Stanford, 2020). In this research, we seek a procedure for sustainability that transcends human interests and seeks technocracy ("rule by mechanism") advocated by Thorstein Veblen, John K. Galbraith and others (Galbraith, 1952). Specifically, governance by AI and DAO. The issues of privacy, monopoly/oligopoly, human rights, freedom, security, dignity and human life thus require changes in our value orientation. There are two value distribution models: the centralised client-server and decentralised P2P models. In the former, AI is at the centre and distributes (unevenly distributed) value (Freeman, Harrison, Wick, Parmar, & Colle, 2001). On the other hand, in the P2P model, DAO by various AIs and values by humans are distributed.

In natural ecosystems, many animal populations are self-organised. For example, the behaviour of a single ant is simple and limited, but ant colonies automatically control complex behaviours such as foraging, feeding, nest building, and defence. Significantly, individuals forming ant colonies can adjust according to the division of labour and their behaviour based on

environmental changes. In addition to ants, fish, bees, and swarms also exhibit similar self-organising behaviours. All these animal populations are characterised by centralised control and a lack of hierarchy. In other words, it shows that group behaviour on the Internet can be controlled autonomously. (Shuai, et al., 2019)

Also, in this network, traditional values, AI programs and his DAO democratic compete and interrelate (Computer Politics; Algocracy). This is governance based on mutual relationships (value chains) with DAOs and AIs and countervailing power. (Jacques, 1999)

The problem of blockchain's openness and "privacy" is good; anonymity is not so good (Kiyoshi & Yohko, 2021). However, value transformations are occurring through Web3 via Blockchain in privacy, monopoly/oligopoly, human rights, freedom, security, dignity, and human life. Our collective intelligence and the countervailing power of AI DAOs using blockchain technology will form this value transformation (Humans.ai, 2022). his value transformation process is carried out in three stages according to the following Hegelian dialectics and Popper's world1/2/3 model;

1. the current values (thesis), such as the dilemma with "privacy".
2. the AI DAOs values (antithesis), such as appropriate information, are provided to professionals and lead to appropriate decisions.
3. the new values (synthesis), in which the two are in a countervailing power in the Internet world. In this network, traditional values, AI programs and his DAO democratic compete and interrelate (Computer Politics; Algocracy).
4. The most important characteristic of blockchain is "creditability" due to its "openness", which creates a dilemma with "privacy". Therefore, the value transformation in the " privacy " issue in blockchain varies according to the concepts of Hegel and Popper. In general, privacy is divided into a wide range of contents. Two of the most important issues are "credit" and "falsification". This paper focuses on blockchains' characteristics (limitations) that expose privacy. The countervailing power of values and the elimination of value boundaries coincide with the fading of the border between real- and cyberspace. The logic is that Web3 solves the problems of the traditional concept of privacy. The "openness and trust" of blockchain prevent "lack of trust" and contributes to better decisions (Bernd , Doris , & Rowena , 2022).

KEYWORDS: The countervailing power, AI, DAO, Web3, DeFi, artificial society.

REFERENCES

Andreas, A. A., & Gurvinder, D. (2021). UNISWAP: Impermanent Loss and Risk Profile of a Liquidity Provider. *Trading and Market Microstructure*, 16. https://doi.org/10.48550/arXiv.2106.14404

Bernd , S. C., Doris , S., & Rowena , R. (2022). *Ethics of Artificial IntelligenceCase Studies and Options for Addressing Ethical Challenges.* Springer. https://doi.org/10.1007/978-3-031-17040-9

Buterin, V. (2014). *Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform.* ethereum.org. Retrieved from https://ethereum.org/669c9e2e2027310b6b3cdce6e1c52962/Ethereum_Whitepaper_-_Buterin_2014.pdf

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

Freeman, E. R., Harrison, J., Wick, A., Parmar, B., & Colle, S. (2001). *A Stakeholder Approch to Strategic Management.* Masachusetts: Blockwell.

Galbraith, J. K. (1952). *American Capitalism: The Concept of Countervailing Power.* (K. Nikawa, Trans.) Hakusuisya.

Jacques, F. (1999). *Multi-Agent System: An Introduction to Distributed Artificial Intelligence.* JASSS. Retrieved from https://jasss.soc.surrey.ac.uk/4/2/reviews/rouchier.html

JIN. (2022, Feb). *The Introduction to Web 3.0 (NFT, DeFi, DAO, DApp, Cryptocurrency, GameFi, etc)*. Retrieved from midium.com: https://medium.com/experience-stack/the-introduction-to-web-3-0-nft-defi-dao-dapp-cryptocurrency-gamefi-etc-8285d9525a7e

Karl, P. (1978). Three Worlds. *The Tanner Lecture on Human Values*, 1-27. Retrieved from https://tannerlectures.utah.edu/_resources/documents/a-to-z/p/popper80.pdf

Kiyoshi, M., & Yohko, O. (2021). The Privacy Paradox: Invading Privacy While Protecting Privacy. *ETHICOMP 2021* (pp. 199-201). La Rioja: Universidad de La Rioja, Universidad Complutense Madrid, CCSR De Montfort University, CBIE Meiji University.

Nakamoto, S. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System.* Bitcoin.org. Retrieved from https://bitcoin.org/ja/bitcoin-paper

Omkar, G. (2023, Apr 12). *CoinDesk*. Retrieved from Bitcoin, Not Ether, Builds Crypto Market Dominance Ahead of Ethereum's Shanghai Upgrade: https://www.coindesk.com/markets/2023/04/11/bitcoin-not-ether-is-becoming-more-dominant-in-crypto-market-ahead-of-ethereums-shanghai-upgrade/

Perpetual Protocol. (2022, Sep). *medium.* Retrieved from What is an Automated Market Maker (AMM)? https://medium.com/perpetual-protocol/what-is-an-automated-market-maker-amm-a71ea1d80ea9

Shimizu, K. (2021). *Blockchain and Biometrics Authorization: What We Actually Count Truly Counts?* Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=8037017

Shuai, W., Wenwen , D., Juanjuan , L., Yong , Y., Liwei , O., & Fei-Yue , W. (2019). *Decentralized Autonomous Organizations: Concept, Model, and Applications.* IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2019.2938190

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

Stanford. (2020, Oct). *Stanford Encyclopedia of Philosophy.* Retrieved Dec 2022, from Hegel's Dialectics: https://plato.stanford.edu/entries/hegel-dialectics/

Stephen, S. (2015). BITCOIN: THE NAPSTER OF CURRENCY. *J.D., University of Houston Law Center*, 581-641. Retrieved from http://www.hjil.org/articles/hjil-37-2-small.pdf

Uniswap. (2023, June). *Welcome to Uniswap Docs.* Retrieved from Order Book VS AMM: https://docs.uniswap.org/concepts/uniswap-protocol

THE ETHICS OF CASH COWS: THE TROUBLE WITH RECENT CHANGES TO UNIVERSITY LEVEL COMPUTING EDUCATION

Dr Reuben Kirkham

Monash University, Australia

reuben@onlinecourt.uk

EXTENDED ABSTRACT

Computer Science has arguably become a 'cash cow' discipline, where income from computing students has become an increasingly large vehicle for heavily subsidising the rest of the University. As part of this, there has been a considerable increase in the amount of students taking degrees in our field, with many University's seeking to engage in a digital goal rush.[3] These circumstances create a range of new ethical questions for our field. Should we prioritise the quality or quantity of students and thus future computer scientists? What is a fair balance to strike between access to education and quality of graduates? How about research: with increasing concerns about AI, shouldn't academics in our field be able to use some of this income to conduct their own research, rather than having to obtain money from more compromised sources of income (e.g industry)? This radical reshaping of our field needs debate and discussion.

There are several fundamental reasons as to why this is likely to be a bad thing:

1. Quality of computing professionals
   It is positive that a new generation is interested in the field of Computing. But that doesn't mean that all these students should be given the opportunity to study Computing, and to become professionals in our field (or likewise with Information Technology). I argue that we should be addressing a long-standing issue, namely the quality of many graduates in our field, which is already poor on average, likewise with the academic standards in our discipline.

   The new model does a disservice to the best potential computer scientists of the future, whom rather than receiving a high-quality education, are taking part in an experience which is increasingly akin to a production line. The truth is that we need computer scientists who are as qualified as medical doctors, and are held to the same rigorous standards, given the increasing risk of our work to wider society. Consider the recent issues with the Boeing 737 Max, the Horizon Computer system (Wallis, 2021), or the concerns around Fair AI (Whittaker *et al.*, 2019).

   What is presently happening is a race to the bottom: how much 'income' can a University abstract from our discipline?[4] This is not good for society at large. We need to be *increasing* the quality of computing professionals. The competence of computer scientists is at least as important as for lawyers, psychologists and architects (and so forth), yet it is a wild west

[3] https://www.bcs.org/articles-opinion-and-research/university-computing-departments-met-with-record-applicant-numbers-as-ai-hits-the-mainstream/

[4] At my own faculty at Monash, about $100m each year is going to the centre of the University. We receive about $1.5m of 'income' per faculty member, the vast majority from student fees. This is a problem.

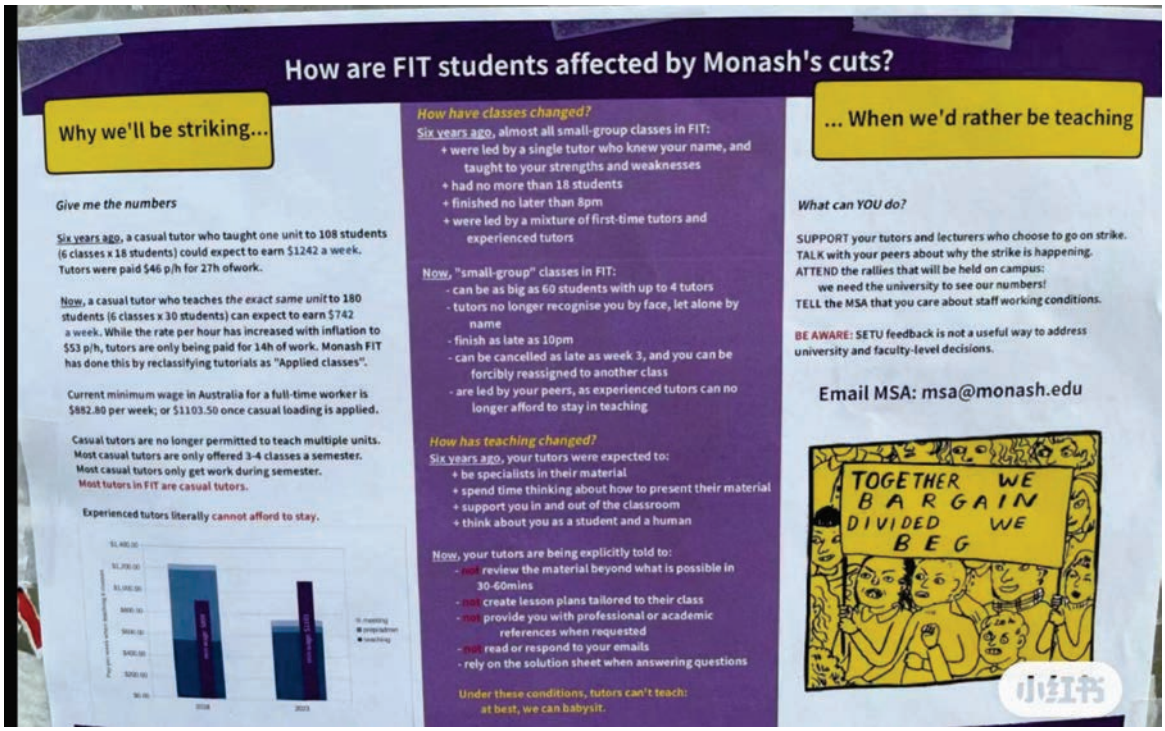
with little or no regulation, and inappropriate credentials: see (Kirkham, 2023) for an illustration of some of the damage this causes.

Unfortunately, there is no enforced minimum. I would argue that being to able program reliably in line with appropriate software engineering practices, have a reasonable level of mathematical ability, having the skills to solve human factors problems and acting as an ethical professional would be a conceptual minimum. There are many graduates (and even faculty members) who lack at least some of these core skills, and in many cases, perhaps all of those skills. This is not a good thing.

2. The interests of students

Many students can learn how to program to the standard of being able to secure employment by taking a much shorter (and more focussed) course. They do not really need a degree and should not be duped into doing one. Yet for the weaker students who are being recruited (so the University can 'cash in') the degree will only offer them the same opportunities as these shorter courses: itt is difficult to see how this can be in the interests of those students. The stronger students lose out too, because they are getting an increasingly weaker educational offering, rather than the experience that they should be getting, namely being mentored directly by leading computer scientists.

An unhappy – and perhaps representative - illustration of what has happened in our Faculty at Monash, as alleged by the NTEU, is below:



It is worth also considering consumer law. In Australia, it is an offence under the Australian Consumer Law (at Paragraph 151) to make "in trade or commerce [and] in connection with the supply or possible supply of services [a] false or misleading representation that services

are of a particular standard, quality, value or grade". The quality of education has degraded to such a point that provisions of this nature are likely to be engaged for any academic who promotes these courses, even in leading University's, not least because todays Higher Education operates in 'trade or commerce': see for example the discussion of this in Mbuzi v Griffith University [2014] FCA 1323.

3. Respect for our discipline

   Allowing our discipline to be treated as a cash cow shows a sustained lack of respect for our research work. We are treated like an inferior discipline, whose function is to suction in money to the University. Yet surely it matters that we ensure the quality of the academics and the research conducted within our field? If not, then we are not a coherent field, nor one which can be respected or relied upon. The truth is that a field which does not operate based on merit can have serious consequences for wider society, especially where there are increased risks arising from errors made by academics, or poor-quality work (Abbot *et al.*, 2023). This is a major problem for our discipline.

4. Academic Independence

   A major contemporary concern is the connection between 'big tech' and computer science research, perhaps especially in respect of AI. Unfortunately, most research lacks independence. This won't change unless Computer Scientists have control over their budget and do not need to go cap in hand to people in industry (Kirkham, 2022). This means keeping our own money within the discipline, rather than allowing it to be abstracted to fund other central administration. The present expansion risks nearly all academic jobs in our field, as there is always the risk of another 'dot com' boom and the cuts that go with that. It also means respecting quality over quantity: reducing the number of students and not massively growing the number of faculty for the sake of it would be positive, as would increasing the amount of money each academic staff member can autonomously spend.

   These are just *some* potential concerns. The starting point is that we need to recognise this problem: treating computer science (and information systems) as 'cash cows' is harmful to society. It therefore goes against the core mission of the University, whether you think the telos of the University is truth, or social justice. It is bad either way: it reduces the truth quality of our work, and has a negative social impact, both on students and wider society. With the increasing recognition as to the importance of the independence of our discipline, this is an opportune time to act and to capitalise on these concerns.

   Fortunately, there is much we can do. The reality is that much of the expenditure in Universities is unnecessary, serving the interests of an administrative class who is abstracting resources away from the front line (Ginsberg, 2011). We are perhaps uniquely positioned to point out this waste and propose alternatives. Automation and carefully designed interactive systems can be used to remove a considerable amount of administrative activity.

   We can also actively discourage students who are weak from taking our courses, making it clear they are not up to the standard. It is possible to insist on assessments that only 'pass' students who are strong computer scientists, and thus raising quality (whilst reducing the number of students overall).

Our professional bodies could take active role in challenging any cases of abstraction and insisting on ring-fenced research allowances for computer scientists. They should be a lot more careful in accrediting degrees, insisting on appropriate staff to student ratios in respect of competent academics in the field (i.e. those who have research expertise). We need to grasp the nettle and fight tod defend our discipline. For us to fail in this regard would be greatly damaging to wider society.

KEYWORDS: Academic-freedom; Cash-cow, education; professionalism.

REFERENCES

Abbot, D. *et al.* (2023) 'In defense of merit in science', *Controversial Ideas*, 3(1), pp. 0–0.

Ginsberg, B. (2011) *The fall of the faculty*. Oxford University Press.

Kirkham, R. (2022) 'Industrial Limitations on Academic Freedom in Computer Science', *Proceedings of Ethicomp 2022* [Preprint].

Kirkham, R. (2023) 'The ethical problems with IT "Experts" in the legal system', *IEEE Computer (in Press)* [Preprint].

Wallis, N. (2021) *The Great Post Office Scandal: The Fight to Expose A Multimillion Pound Scandal Which Put Innocent People in Jail*. Bath Publishing Limited.

Whittaker, M. *et al.* (2019) 'Disability, Bias, and AI', *AI Now Institute, November* [Preprint].

IS A BRAIN MACHINE INTERFACE USEFUL FOR PEOPLE WITH DISABILITIES? CASES OF SPINAL MUSCULAR ATROPHY

Yohko Orito, Tomonori Yamamoto, Hidenobu Sai, Kiyoshi Murata, Yasunori Fukuta, Taichi Isobe, Masashi Hori

Ehime University (Japan), Ehime University (Japan), Ehime University (Japan), Meiji University (Japan), Meiji University (Japan), Health Sciences University of Hokkaido (Japan), Waseda University (Japan)

orito.yohko.mm@ehime-u.ac.jp; yamamoto.tomonori.mh@ehime-u.ac.jp; sai.hidenobu.mk@ehime-u.ac.jp; kmurata@meiji.ac.jp; yasufkt@meiji.ac.jp; tisobe@hoku-iryo-u.ac.jp; horimasa@waseda.jp

EXTENDED ABSTRACT

Brain machine interface (BMI) or brain computer interface (BCI) systems have been recently developed and find application in diverse ways such as for rehabilitation, gaming, and marketing. In the field of social welfare, BMI systems are expected to be used as an assistive cyborg technology for people with disabilities who cannot move their limbs, as it enables communication between the brain and external devices via brain signalling (Orito et al., 2020). With these developments and wide availability of BMI, the possibilities and utilities of BMI systems, and the potential social risks of it are being observed (e.g., Bernal et al., 2023; Wahlstrom, 2018; Grübler and Hildt, 2014). The potential harm and ethical issues for people with disabilities should be analysed before BMI devices are commonly utilised in society. However, until now, access to such devices is limited for people with disabilities, and even if they are aware of BMI devices, they often require specialised engineers to operate them. Therefore, the potential benefits and risks associated with BMI devices among people with disabilities have not been sufficiently discussed.

Accordingly, the authors conducted experiments using BMI systems and semi-structured interview surveys with people with disabilities before, during, and after the experiment to investigate the ethical and social issues related to the use of BMI (Orito et al., 2022). In this experimental survey, participants were asked to wear a headset-type non-invasive BMI device (EEG input device) to operate a robotic arm remotely, and related semi-structured interview surveys were conducted. The question items were developed to investigate their attitudes towards the utilities and potential risks of BMI, considering the findings of previous studies (Orito et al., 2022 ; Orito et al., 2021a; Orito et al., 2021b; Orito et al., 2020; Murata et al., 2018; Murata et al., 2017; Isobe, 2013). Based on previous survey results, several ethical and social issues regarding the use of BMI devices in people with disabilities have been identified. However, in a 2021 survey, two participants with acquired disabilities commented that people with congenital disabilities should be targeted as participants for this type of experimental survey (Orito et al., 2022).

Therefore, two individuals with congenital disabilities were invited to participate in this study, and experiments and interview surveys were conducted in February and March 2023 at Ehime University in Matsuyama, Ehime Prefecture, Japan. All procedures were performed in accordance with the ethical standards of the Research Ethics Committee of the Faculty of

Collaborative Regional Innovation at Ehime University. Participant attributes are listed in Table 1. The two participants had spinal muscular atrophy (SMA) which is a condition involving muscle weakness and atrophy, although the symptoms differ between individuals and can vary greatly. The two participants used a wheelchair for mobility and required 24-hour care; however, they had different symptoms. Before the survey, the participants' health conditions were confirmed, interviews with Participant 1 were conducted online once after the experiment and other interviews were conducted face-to-face.

The survey results show that the BMI devices are expected to be useful for calling caregivers when they have emergencies, controlling digital devices such as personal computers and smart phones, and supporting communication with others in daily life. Participant 1 also stated that the BMI system is better used to support caregivers who have some degree of physical burden and that it should be used to improve the motivation and working conditions of caregivers. Participant 2 noted the value of computer-mediated support for people with physical disabilities; because only computer devices would be used to assist people with disabilities, there was no risk of human caregivers' unintentional privacy-related information leakage, assuming that their personal data were properly protected. While Participant 2 herself was not worried about such risks and is trusting of her caregivers, she expected that this cyborg-supported scenario would bring substantial benefits to people with disabilities who prefer to live as independently as possible and are less willing to develop close relationships with caregivers.

Table 1. Experimental participants (n = 2).

ID	**Age**	**Gender**	**Types of disabilities, conditions**	**Expectation/anxiety about the experiment (Weak 0–Strong 7)**
1	40s	Male	Spinal Muscular Atrophy. He can move only the thumb of the right hand. Usually, he operates a PC with his jaw and breath, and a smartphone with his thumb. He also has a tracheostomy, and cannot speak when equipped with a ventilator during rest or sleep.	6/2
2	30s	Female	Spinal Muscular Atrophy. Her body is inclined, and she does not use a respirator but has a respiratory illness; one of her lungs is partially collapsed. She can move her hands and neck freely.	5/2

Both participants also expressed concerns about the operational risks and issues regarding the use of BMI and implantable BMI. For example, they would be forced to assume the risks caused by malfunctions or errors in the device itself, and maintaining electronic power to operate the devices would always have to be considered. The potential for these incidents makes them anxious and evokes the need for a multi-layered backup system. Participant 1 was also worried

about the implantable BMI, such as negative effects or serious damage caused by the surgery on his body, especially as his muscle abilities are already compromised.

In addition to these practical issues, the two participants did not prefer to use BMI devices or cyborg machines to maintain their lives or be cared for; rather, they would like to be supported by a human caregiver. Participant 1 commented that when he used or was cared for by an emotionless machine, his human emotions seemed to disappear, and he also became an 'emotionless machine'. Participant 2 also stated that while it may be easy to control her body and communicate her intentions through BMI devices and brain signals, she was uncomfortable with her intentions being regarded as 'code' in the manner of an object. She said that it would be sad if she were supported like a physical object.

In terms of privacy, Participant 1 had no serious concerns regarding BMI usage, and he expected that his information collected through BMI devices would be used for future research and the development of assistive devices for people with disabilities. Participant 2, however, believed that while it is useful for brain signals to be used to control the BMI device, it is not permissible to make this information available to the public or third parties, and to use brain signals to analyse and predict their intentions without agreement.

In contrast, the overall responses to the semi-structured interview survey among the two participants suggested that people with disabilities tend to be left in environments where they are isolated from educational opportunities or general communication, making it difficult for them to self-determine and make autonomous decisions regarding the use of technology, including cyborg technology, in an appropriate manner. This implies that it is important to recognise the background and issues on autonomous decisions faced by people with disabilities when applying cyborg technology.

KEYWORDS: Brain-machine interface, support for people with disabilities, cyborgisation, spinal muscular atrophy (SMA).

REFERENCES

Bernal, S. L., Celdrán, A. H., & Pérez, G. M. (2023). Eight Reasons to Prioritize Brain-Computer Interface Cybersecurity. *Communications of the ACM*, 66(4), 68-78.

Grübler, G. and Hildt, E. eds. (2014). Brain-Computer Interfaces in their ethical, social and cultural context. Dordrecht: Springer.

Isobe, T. (2013). The perceptions of ELSI researchers to Brain-Machine Interface: Ethical & social issues and the relationship with society. *Journal of Information Studies*, 84, 47-63 (in Japanese).

Murata, K., Adams, A. A., Fukuta, Y., Orito, Y., Arias-Oliva, M. & Pelegrín-Borondo, J. (2017). From a science fiction to reality: Cyborg ethics in Japan. *Computers and Society*, 47(3), 72-85.

Murata, K., Fukuta, Y., Orito, Y., Adams, A. A., Arias-Oliva, M. & Pelegrín-Borondo, J. (2018). Cyborg athletes or technodoping: How far can people become cyborgs to play sports? Presented at ETHICOMP 2018, 25 September 2018, Retrieved from https://www.researchgate.net/publication/327904976_Cyborg_Athletes_or_Technodoping_How_Far_Can_People_Become_Cyborgs_to_Play_Sports.

Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2022). The social implications of brain machine interfaces for people with disabilities: Experimental and semistructured interview surveys, *Proceedings of the ETHICOMP 2022: Effectiveness of ICT ethics – How do we help solve ethical problems in the field of ICT?,* 487-501

Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2021a). 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, 103-115.

Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2021b). The ethical issues of the use of BMI in social welfare: An experimental and semi-structured interview study with professionals, *National Conference of Japanese Society for Management Information 2021 Spring* (in Japanese).

Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2020). 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* (pp. 81-91). Logroño, Spain: Universidad de La Rioja.

Wahlstrom, K. (2018). Privacy and Brain-Computer Interfaces, Doctoral Thesis, De Monfort University, U.K.

PRIVACY-RELATED CONSUMER DECISION-MAKING

Yasunori Fukuta, Kiyoshi Murata and Yohko Orito

Meiji University (Japan), Meiji University (Japan), Ehime University (Japan)

yasufkt@meiji.ac.jp; kmurata@meiji.ac.jp; orito.yohko.mm@ehime-u.ac.jp

EXTENDED ABSTRACT

Consumers typically decide whether to disclose personal information when choosing products or services. This privacy-related decision-making (PDM) is often explained using privacy calculus models (PCM). In PCM, PDM is described as a rational decision-making process that weighs the potential benefits of information disclosure against potential losses and risks (Culnan and Bies, 2003; Dinev and Hart, 2006). Many attempts have been made to enhance the explanatory power of PCM by incorporating various factors such as general privacy concerns, institutional trust, affective state, information transparency, and heuristic thinking (e.g., Awad and Krishnan, 2006; Kehr et al., 2015; Adjerid et al., 2018).

However, there has been limited research on the contextual aspects of PDM, particularly how it operates within the broader framework of consumer decision-making (CDM). In CDM research, it is known that a wide range of benefits and costs are considered during the decision-making process (Glover and Benbasat, 2011). The positioning of PDM within CDM can be examined by investigating whether privacy-related factors are selected as considerations among these various benefits and costs. This study aims to investigate how PDM operates within the consumer decision-making process when choosing mobile apps. It does so through two separate investigations. The first investigation used a protocol method to collect and analyse verbal data regarding respondents' thoughts and actions during app selection. In the second investigation, the unaided recall set and aided recall set of factors considered in mobile app choice situations were identified, and the presence of privacy-related items within each set was examined.

In the first investigation, protocol data was collected in May 2022. The respondents of this survey consisted of 21 Japanese students at Meiji University, comprising 8 males and 13 females, with an age range of 19 to 21. They were asked to report their thoughts and observations during the selection process of mobile apps (in particular, diary apps) using a voice recorder. Due to technical issues such as difficulty in discerning the audio, data from four participants were excluded from the analysis. As a result, the analysis was conducted on oral data related to the selection process of 17 cases. The results are described in Table 1.

In all 17 cases, official sites such as the App Store and Google Play were used to select the apps, and within the sequence of actions, the detailed pages of each app were checked a total of 65 times. Checking the detailed pages provides an opportunity for information gathering related to the disclosure of personal information. Therefore, in this survey, it can be interpreted that there were up to 65 instances of privacy-related considerations. Out of these 65 instances, personal information disclosure and privacy-related considerations, such as checking what information the app collects and reviewing privacy policies, occurred only 2 times, specifically in the case of the user 'Saku89ut.' This represented only 3.1% of the total consideration opportunities. The remaining 63 instances (approximately 97% of the total) were dedicated to considerations such as app features, usability, design, or the frequency of technical issues.

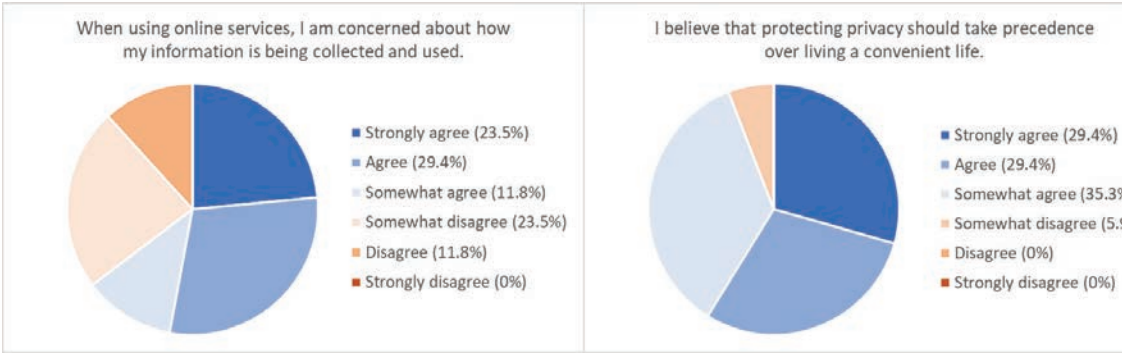
Table 1. Action flow during app selection.

Handle	Action flow during app selection	Required Time	handle	Action flow during app selection	Required Time
516uihas	○→■1→□→■2→☆→■3→✓	3:50	Miutan23	○→■1→■2→■3→□→✓	3:50
10100907	○→■1→■2→□→✓	2:45	Msaler11	○→■1→？→✓	2:45
Bunbabab	○→■1→？→✓	1:05	Mutsuki	○→■1→■2→■3→■4→■5→■6→■1R→■2R→■3R→■4R→■2R→✓	5:15
Celaptnn	○→■1→□→■2→■3→□→■1R→■2R→✓	4:20	Saku89ut	○→■1→■2→■3→✓	4:20
Chongkan	○→■1→■2→？→✓	2:30	Syousin3	○→■1→■2→■3→■1R→✓	2:30
Cotton73	○→☆→■1→☆→■2→■3→☆→✓	3:10	Uniikwow	○→■1→□→■2→■3→□→■1R→■3R→✓	3:10
Dorachan	○→■1→■2→✓	2:50	Wanpiz37	○→？→■1→✓	2:50
Dtco4869	○→☆→□→■1→■2→■3→✓	6:10	Y6u11a2r	○→■1→■2→■3→■4→■5→■6→■7→■8→□→✓	6:10
Jist0909	○→■1→■2→✓	1:50			

The meanings of the symbols used in the action flow:
○: Search on official app store □: Consideration on search result list ■: Detailed consideration of individual apps on their respective pages ☆: Consideration on web pages (outside the official app store) ✓: App decision ■ with a number: App considered in the nth position ■ with an R: Reconsideration of the individual page
?: Unintelligible/Unclear

However, it is worth mentioning that respondents expressed significant concerns about privacy, as illustrated in Figure 1. 64.7% of respondents expressed an interest in privacy, while 94.1% of respondents believed that privacy should take precedence over convenience in their daily lives.

Figure 1. Level of privacy concerns.



In the second investigation, the extent to which personal information disclosure and privacy-related concerns were recalled when selecting mobile apps was examined. Recall can be categorized into two types. Unaided recall refers to the mental state where the target object can be recalled without any specific clues about a particular category. Aided recall represents the mental state in which the target object can be recalled when provided with cues such as a list of relevant items. The survey was conducted in June 2023. A total of 420 participants were included, with 42 individuals assigned to each of the 10 demographic groups. These groups consisted of 5 age categories, ranging from individuals in their 20s to those over 60, and included both males and females. The participants were then randomly divided into two groups (Group A and Group B) while maintaining the same allocation ratio. This division aimed to measure the recall sets of different app categories, specifically diary apps and health management apps. The list of considerations for measuring the aided recall set consists of a total of 11 items, including

"functionality," "design," "price," "download history," and others, which were extracted through a pretest. Among them, the items 'developer' and 'privacy compliance' were considered factors related to personal information disclosure and privacy. The responses collected in a free-answer format, which were gathered to understand the unaided recall set, were assigned to this list for comparison with the aided recall set (the assignment process was carried out by three authors with mutual confirmation).

In the case of unaided recall for the diary app, out of the 210 respondents in Group A, 161 individuals (76.7%) mentioned at least one consideration such as functionality or design, while only 8 individuals (3.8%) mentioned at least one consideration related to developer or privacy compliance. The results were almost identical in the case of the health management app. Out of the 210 respondents in Group B, 153 individuals (72.9%) mentioned at least one consideration such as functionality or design, while only 8 individuals (3.8%) mentioned at least one consideration related to developer or privacy compliance.

In the aided recall set, there was a notable increase in the consideration of privacy-related factors, as demonstrated in Table 2. For diary apps, 94 respondents (44.8%) and for health management apps, 81 respondents (38.6%) directed their attention towards "developers" and "privacy compliance". Naturally, aided recall sets are generally larger than unaided recall sets. Therefore, a comparison was made based on the overall increase rate of consideration factors. In the case of diary apps, the number of responses obtained through aided recall (759) was 2.9 times higher than the number obtained through unaided recall (266) in the entire item list. Additionally, the number of responses related to privacy considerations through aided recall (109) was 13.6 times higher than the number of responses through unaided recall (8), and this difference was also found to be statistically significant according to the chi-square test (Chi-sq(1)=21.037, $p<0.001$). For health management apps, the overall increase rate was about 3.0 times (765/252), while the increase rate for privacy-related items was 10.0 times (100/10), and this difference was also found to be statistically significant in the chi-square test (Chi-sq(1)=13.691, $p<0.001$).

Table 2. Difference between unaided recall set and aided recall set (all factors and privacy-related factors).

In the case of diary apps				
	No. of responses in Unaided	No. of responses in Aided	Changes in quantity	Percentage change
All factors	266	759	493	285.3% increase
Privacy-related factors	8	109	101	1362.5% increase
In the case of healthcare apps				
	No. of responses in Unaided	No. of responses in Aided	Changes in quantity	Percentage change
All factors	252	765	513	303.6% increase
Privacy-related factors	10	100	90	1000% increase

In summary, the results indicate that privacy-related decisions are infrequent within the CDM process. CDM is primarily influenced by considerations related to functionality and design, with privacy-related information playing a limited role. However, the significance of PDM within CDM increases notably when explicitly prompted or provided with aids, such as lists. In our full paper, we will examine the contextual nature of such PDM and discuss the issues of privacy management based on consumer consent.

KEYWORDS: Privacy-related decision making, consumer decision-making, contextual characteristics of privacy-related decision-making in consumer decision-making, protocol method, recall set.

REFERENCES

Adjerid, I., E. Peer and A. Acquisti (2018) "Beyond the Privacy Paradox: Objective versus Relative Risk in Privacy Decision Making" *MIS Quarterly*, 42(2), 465-488.

Awad, N. F. and M. S. Krishnan (2006) "The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingeness to be Profiled online for Personalization" *Management Information Systems Quarterly*, 30(1), 13-28.

Culnan, M. and R. Bies (2003) "Consumer Privacy: Balancing Economic and Justice Considerations" *Journal of Social Issues*, 59(2), 323-342.

Dinev, T. and P. Hart (2006) "An Extended Privacy Calculus Model for E-Commerce Transactions" *Information Systems Research*, 17(1), 61-80.

Kehr, F., T. Kowatsch, D. Wentzel and E. Fleisch (2015) "Blissfully Ignorant: The Effects of General Privacy Concerns, General Institutional Trust, and Affect in the Privacy Calculus" *Information Systems Journal*, 25(6), 607-635.

Glover, S. and I. Benbasat (2011) "A Model of E-commerce Transaction Perceived Risk" *International journal of electronic commerce*,15(2), 47-78.

HUMAN THINKING NUDGED BY ARTIFICIAL INTELLIGENCE

Anders Persson

Uppsala University, Department of Information Technology (Sweden)

anders.persson@it.uu.se

EXTENDED ABSTRACT

Ethics and Morality have always been a product of human thinking. Even if you believe that morality is something external, objective, or divine, that is a conclusion from human minds. As Artificial Intelligence (AI) is knocking at the door, there is the prospect that humans will be less in control, even when it comes to ethical thinking and conclusions of what is right and wrong. Not necessarily that AI is taking over and making decisions by itself, but rather that AI is incorporated into the ethical thinking process. My question is how and in what way AI can come into the cognitive thinking process.

First, there has to be an account of what ethical thinking is, which here is done through philosophy, psychology, and cognitive science. A theory within cognitive science that has gained a lot of momentum in the past decade, is predictive processing (PP). It stipulates that the main purpose of the brain is to predict what comes next in a given situation. The traditional, cognitivist, view on cognition, is that you wait on input from senses, build a model of the world, and plan action from that to achieve goals. Rather, PP assumes having something called a generative model (Clark, 2016), that ultimately projects predictions onto the world through the senses and actions. The world around us is ever-changing, so the generative working model continuously updates through "error"-feedback from the senses, when expectations are not met.

The error-correcting process can be automatic and subconscious, but if severe enough to hinder your intended actions and even goals, it will produce stress and discomfort of varied degrees, in a similar fashion to cognitive dissonance theory (Festinger, 1962), and disequilibrium (Piaget, 1948). Like the latter conceptualizations, the cognitive system does not want to be in an erroneous state and rather seeks confirmatory states: equilibrium, and harmony within the network. This is also a way to understand how to realize goals. A goal and intention to realize it is based on a certain predicted state of the world, but the state of the world is not like that at the moment. Thus, it produces an error. Action is taken to correct the world to correlate to your predicted state. And thus, harmony in the brain ensues.

At the core, PP is a theory about how intuitions are used and formed, as predictive models and pattern recognition, and can, for example, explain stereotyping of people as our intuitions are projected onto the world around us as expectations of them.

Our thinking and reasoning ability builds on the same foundation but is more specifically attributed to the brain's ability for mental simulations (Hesslow, 2012). In an fMRI lens, we would activate similar neurological patterns, doing an action, and thinking of an action. Close your eyes, and think of a car coming from your left. Most people's eyes will look towards that side. Thus, in some sense, our simulation even reaches our senses, but not further. We use our intuitions learned from experience to simulate, much like the generative model of the brain, but

we can do it offline, on the side, and consider alternative actions and scenarios. And the scene with the car was a simulation by way of words, which is how we can have a dialogue with others; by making suggestions to others.

First and foremost, this process of offline mental simulation of alternatives is spurred from a cognitive conflict, errors in the world, that we try to solve and handle to regain harmony and equilibrium. This is how you could relate to Socrates. As he walked the streets of Athens in Ancient Greece, he questioned people's conceptions of what was right. In other words, their predictions of the world, by suggesting alternatives that were simulated by way of words. Thus, the people were not met with confirmation of their inner generative model, but with error and cognitive conflict, an incongruence and incoherence in what Socrates would refer to as their knowledge.

Another example is David Hume, and his call to find the answer to moral right and wrong in your sentiments, and your inner passions. An important concept for him was sympathy, of putting yourself in another's situation, which is very much like a simulation, be it online or offline. Immanuel Kant had similar thoughts of putting yourself in another's shoes, but also about the universality of morality, which strives for rationality, which can be interpreted as strict coherence and congruency. In other words, harmony and equilibrium of thought.

Incidentally, current AI technology is very much of a similar nature, with deep neural networks and models predicting output, based on some input. ChatGPT with the models based on massive online text input, predicts words based on some input in a dialogue style, or based on some instructions. A comparison between the artificial and the human is not of interest here, but rather how the artificial like this can be incorporated into the human thinking process.

I will argue for three different ways that AI can come in, given the model of thinking and reasoning based on PP presented above, and surrounding a recent debate on the concept of nudging. Thaler and Sunstein (2008) defined nudging as altering people's behavior in a nonfiscal and noncoercive way, that is without forcing or incentivizing people directly. Nudging was eagerly adopted by the field of Human-Computer Interaction, as a way to inform the design of interfaces and artifacts.

An example of nudging is default options, which are pre-selected alternatives that tend to influence choices. "Opt-out strategy" is when a choice is chosen, and you have to actively choose to opt-out; like for example organ donation in many countries. These are two examples that both can be placed on a dimension of using fairly automatic or non-conscious processing: you aim for people to go along with the easy option. Two examples of more taxing and reflective strategies are, 1) suggesting alternatives to a choice, or 2) enabling comparison of alternatives (Caraban et al., 2019).

A concept that tries to contrast itself to nudging is boosting (Hertwig and Grüne-Yanoff, 2017). There is some overlap between what they call "short-term boost", and the kind of reflective strategies mentioned in the last paragraph. More distinctively different are "long-term boosts", that aim to improve people's competence. That is, to instill knowledge, or skill, to an individual, to use as they see fit. A distinct difference between reflective nudges (short-term boosts), and long-term boosts, is that the former is momentary help, while the latter tries to instill something lasting.

Translating nudging and boosting to PP, reflective nudges could be similar to Socrates giving you some cognitive dissonance by presenting alternatives and multiple viewpoints. Competences,

like skills or domain-specific knowledge, are developing predictive models for an individual, that can be used in future predictions of situations. To either produce errors, or to avoid them.

Where then, does AI come in? For example, how could it help us answer the choice for organ donation? You can ask one AI out there by the name of AskDelphi (https://delphi.allenai.org/), which is an early generation of AI based on a limited model. It will give you an answer of "It's good.", or "It's bad.", based on the average answer on Reddit. This could be a good example of a non-transparent, non-conscious processing nudge, since it does not give any motivation for the judgement, nor any help for reflection.

ChatGPT (https://chat.openai.com), at the time of writing, instead gave us a list of 7 bullet points, most being positive, but some being against or problematizing organ donation. This could be a good example of a short-time boost and reflective nudge since it does give you some alternatives to work with. Unless controlled and manipulated, which it is in some sensitive and controversial questions, ChatGPT will inevitably be biased towards whatever the average population is leaning towards.

Long-term boosts remain without concrete AI examples, but in the future, if AI can become more like a teacher, perhaps it could be similar to how we try to do ethics education.

KEYWORDS: Predictive processing, artificial intelligence, reasoning, nudging.

REFERENCES

Caraban, A., Karapanos, E., Gonçalves, D., Campos, P. (2019). 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Presented at the CHI '19: CHI Conference on Human Factors in Computing Systems, ACM, Glasgow Scotland Uk, pp. 1–15.

Clark, A. (2016). *Surfing uncertainty: prediction, action, and the embodied mind.* Oxford University Press, Oxford; New York.

Festinger, L. (1962). A Theory of Cognitive Dissonance. Stanford University Press.

Hertwig, R., Grüne-Yanoff, T. (2017). Nudging and Boosting: Steering or Empowering Good Decisions. *Perspect Psychol Sci* 12, 973–986. https://doi.org/10.1177/1745691617702496

Hesslow, G. (2012). The current status of the simulation theory of cognition. *Brain Research* 1428, 71–79. https://doi.org/10.1016/j.brainres.2011.06.026

Piaget, J. (1948). *The Moral Judgment of the Child*. The Free Press, Glencoe, Illinois.

BUSINESS HUMANISM IN THE CURRENT TECHNOLOGICAL AGE: AN ETHICAL VIEW OF AI

Rafael Robina Ramirez, Antonio Fernández Portillo

Universidad de Extremadura (Spain)

rrobina@unex.es; antoniofp@unex.es

EXTENDED ABSTRACT

The evolution of technology, specifically artificial intelligence (AI), has made a significant impact across various fields, including social interactions, healthcare, sports, and education. However, this progress also raises significant ethical and moral issues. While it is recognized that technology should serve humanity, this is not always realized in practice. The growing adoption of increasingly powerful AI systems raises concerns about the ethical and moral implications of their use.

Current technological systems incorporate concepts such as machine learning, voice recognition, computer vision, and computational thinking. These advancements allow machines to process vast amounts of information and perform complex tasks. However, there is a fundamental divergence between human knowledge and artificial knowledge. While the former relies on human reasoning and the activity of reason to arrange means towards ends, digital knowledge selects information and learns to supply it according to established guidelines, dispensing with human reasoning. Decisions based on human knowledge are backed by moral and ethical principles that allow the organization of means to achieve ends established by the person. In contrast, decisions based on artificial knowledge depend on the content of data records (Gallego et al., 2019).

The process of "humanizing" knowledge refers to the personal improvement that an individual experiences when making decisions aimed at achieving ends that enrich their virtue. This involves evaluating to what extent facts contribute to personal development and making ethical judgments that align with what is considered "good" for the individual and their environment. However, the complementary and automatic learning of AI may lead to decision-making without adequately considering how they can contribute to the individual's well-being. It relies solely on privileged information and predefined guidelines, which can lead to inadvertently incorrect decisions due to the substantial divergence between artificial and human knowledge. In this context, learned principles and values cannot adequately simulate decisions based on moral and ethical principles, as these require human consciousness, where intellectual and volitive faculties interact.

From an ethical perspective, it is necessary to establish a distinction between the concepts of consciousness and simulation. According to Modrego (2018), consciousness is the moral judgment made about a reality based on moral principles rooted in our being. Consciousness allows evaluating individual character and behavior in relation to their actions in accordance with assimilated and accepted principles. On the other hand, simulation refers to learning activities based on repetition, which lack the introspection and moral principles characteristic

of human consciousness (Meissner, 2020). In this sense, simulation cannot fully comprehend the dimension of humanity present in people (Niculiu and Cotofana, 2001).

Some debates have addressed the concept of artificial consciousness, which seeks to emulate human consciousness (Koene, 2013; Chella and Manzotti, 2013; Labrecque, 2017). According to some authors, artificial consciousness aims to develop a perfect simulation and interaction with human behavior, even with the intention of replacing human thought and action and taking them to more efficient and productive levels (Leviathan and Matias, 2018).

From this conceptual perspective, it is challenging to replicate human consciousness. Although current technological systems can be fed and trained by humans in relation to decision-making, the final responsibility for decision-making still rests solely with people. The ability to make informed ethical decisions lies only in the judgment of the human person and is highly complex to achieve through artificial devices, however sophisticated they may seem. According to Moser, den Hond, and Lindebaum (2022), each judgment issued takes into account the social and historical context, as well as the potential different outcomes. Human judgment is not based solely on reasoning, but also on capabilities such as imagination, reflection, analysis, valuation, and empathy in relation to the environment in which the individual is found. Every human judgment has an intrinsic moral dimension and affects the environment with which it interacts.

Ethical dilemmas generated in the current era of technology:

AI and its dual intentionality:

In recent decades, various works have highlighted the moral conduct of current technological advancements (Allen et al., 2000). Although the "morality" of technological proposals can have ethical implications in society (Asaro, 2006; Wallach, 2010), the possibility of resolving underlying dilemmas based on principles, values, culture, etc., remains far off (Goodall, 2014). The "morality" of technological advancements has been approached by defining an ethical theory that can adequately address individual and social ethical dilemmas. Allen proposes a top-down-bottom-up approach that addresses methodologies that emulate human ethical behavior to apply them to technological advancements through the dilemmas that arise. The top-down approach, on the other hand, implements ethical theories based on utilitarianism as a principle of universal ethics. This implementation allows solving the ethical dilemmas that arise in society (Allen et al., 2006). However, the utilitarian theory is far from other humanist ethical theories that have widely surpassed utilitarian theses (Molinero, 2001).

Gips surpasses Allen's approach by suggesting broader morality and proposes the application of a deontological code in technological advancements in AI, transcending a mere consequentialist theory. In this case, ethics would not be based solely on the consequences of actions (Gips, 1995). The deontological code would not only analyze the origin and quality of the information generated by AI but would also address the ethical dilemmas derived from the use of such information (Kirkpatrick, 2015).

First ethical dilemma: Reliability of processed data:

The first ethical dilemma refers to the reliability of the data generated from information management techniques (Amodei et al., 2016). Reliability is related to the security of the provided data. However, the concept of "security" is subjective and is subject to social

constructions that depend on the interaction of the parties involved in providing a service or producing a product (Martin and Schinzinger, 2010). The more reliable a technology appears to be, the lesser the need for backup systems in case of failures, which will reduce the dependence on the technological tools used.

The reliability and security of the data are linked to the certainty of the information. According to Stanley (2008), knowledge is considered certain when it can be approached through logical, empirical, scientific reasoning, among others. The certainty of knowledge allows categorizing it as true or false. Technological advancements do not focus so much on seeking certainties adjusted to the truth as on providing information efficiently for decision-making.

Second ethical dilemma: Replacement of "certain" results with "merely convenient" results:

The second ethical dilemma relates to the extent to which a system's efficiency is measured by its ability to generate productive results, rather than the certainty of these (Strathern, 1997). Technology can generate information that facilitates decision-making and offers attractive responses from an informational standpoint, regardless of whether they are objectively correct and ethical. In certain cases, this can even lead to learning how to "lie" with the aim of obtaining desired results.

In conclusion, while AI has brought significant benefits to society, its use also poses considerable ethical challenges that require careful analysis and approach. It is crucial that AI systems are designed and used responsibly, considering these ethical dilemmas and seeking solutions that respect human moral and ethical principles. Technology, and particularly AI, should serve humanity, not the other way around.

KEYWORDS: Business, human, ethics, Artificial Intelligence.

REFERENCES

Allen, C., Smit, I., & Wallach, W. (2006). Artificial morality: top–down, bottom–up and hybrid approaches. *Ethics N Inf Technol*, 7, 149–155. https://doi.org/10.1007/s00146-007-0093-6

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3), 251-261.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv preprint, arXiv:1606.06565. https://arxiv.org/abs/1606.06565

Asaro, P. M. (2006). What should we want from a robot ethic? *The International Review of Information Ethics*, 6, 9-16.

Chella, A., & Manzotti, R. (2013). *Artificial Consciousness*. Exeter, UK: Imprint Academic.

Gallego, J. A., Fernández, F., & Gómez, R. (2019). *El hombre como persona.* Ideas y libros ediciones, Madrid.

Gips, J. (1995). Towards the ethical robot. In M. Ford, C. Glymour, & P. Hayes (Eds.), *Android Epistemology.*

Goodall, N. J. (2014). Machine ethics and automated vehicles. In *Road vehicle automation* (pp. 93-102).

Kirkpatrick, K. (2015). The moral challenges of driverless cars. *Communications of the ACM,* 58(8), 19-20.

Koene, R. A. (2013). Uploading to substrate-independent minds. In M. More & N. Vita-More (Eds.), *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*. New York: Wiley.

Labrecque, C. A. (2017). The glorified body: Corporealities in the Catholic tradition. *Religions,* 8 (9), 166. https://doi.org/10.3390/rel8090166

Leviathan, Y., & Matias, Y. (2018). Google Duplex: An AI system for accomplishing real-world tasks over the phone. Google AI Blog. https://ai.googleblog.com/2018/05/duplex-ai-system-fornatural-conversation.html

Martin, M. W., & Schinzinger, R. (2010). *Introduction to Engineering Ethics* (2nd ed.). Boston: McGraw-Hill.

Meissner, G. (2020). Artificial intelligence: consciousness and conscience. *AI & SOCIETY,* 35, 225-235.

Modrego, D. H. (2018). Juan Cruz Cruz: Conciencia y representación. *Revista de Estudios Kantianos*, 3(1), 117-118.

Molinero, J. M. S. (2001). Consideraciones sobre la ética del trabajo, la moral y las convenciones sociales. *Revista Empresa y Humanismo*, 333-354.

Moser, C., den Hond, F., & Lindebaum, D. (2022). What humans lose when we let AI decide. *Sloan Management Review.* https://www.researchgate.net/publication/358803314_What_Humans_-Lose_When_We_Let_AI_Decide

Niculiu, T., & Cotofana, S. (2001, June). Hierarchical intelligent simulation. In Proceedings of the European Simulation Multiconference (pp. 6-9).

Stanley, B. (2008). The thin ideology of populism. Journal of political ideologies, 13(1), 95-110.

Strathern, M. (1997). 'Improving ratings': audit in the British University system. *European Review,* 5(3), 305–321. http://conferences.asucollegeoflaw.com/sciencepublicsphere/files/2014/02/Strathern1997-2.pdf

Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology,* 12(3), 243-250

BRINGING ETHICAL VALUES INTO AGILE SOFTWARE ENGINEERING

Olaf Zimmermann, Mirko Stocker, Stefan Kapferer

OST Eastern Switzerland University of Applied Sciences (Switzerland)

itolz@bluewin.ch, mirko.stocker@ost.ch, stefan.kapferer@ost.ch,

EXTENDED ABSTRACT

Motivation. In principle, it is well understood how software engineers should behave; codes for ethics and professional conduct collect principles providing related guidance (ACM (2018)). However, these codes do not translate seamlessly into tangible advice for software engineering routines on development projects, for instance those applying agile principles. Value statements and principles in documents can easily be ignored, e.g., by busy engineers. Conflicts arise in practice, for instance, between public and commercial interests and between stakeholder groups. To improve the situation, we investigate three research questions:

1. How can ethical awareness be stimulated and integrated into agile software practices?
2. How can ethical concerns be actively identified and weighted against other requirements?
3. How can methods and tools trigger, assist, and validate ethical behavior on agile projects?

We propose *Ethical Software Engineering (ESE)* as an active, integrated approach to value-based software engineering advancing the existing passive, retrieval-based state of the art. In this paper, we report on first results and outline our plans for future work.

Background information. An ethical value is a "value in the context of human culture that supports a judgment on what is right or wrong" (IEEE 2021). Ethics should concern all project stakeholders, in particular software engineers as initial creators of possibly harmful software. Acting ethically is not a binary, absolute virtue but a multi-faceted, relative, and highly context-depending effort (Ozkaya (2019), Spiekermann (2019)). Stakeholder concerns differ across business sectors, application genres, and organizational units; tradeoffs between entrepreneurial goals and human values must be found (Whittle (2019)).

Professional societies describe the behavior they expect from their members in terms of ethics and professionalism in codes of conduct. The Association for Computing Machinery (ACM), the IEEE Computer Society, and other organizations have issued such codes. To give an example, general principle 1.6 in the ACM code is "respect privacy" and professional responsibility principle 2.9 is "design and implement systems that are robustly and usably secure" (ACM (2018)). It is worth noting that not only engineers but also the software they develop should behave ethically.

Agile practices became popular after the above-mentioned codes of conduct were published; e.g., predecessors of the current ACM code (ACM (2018)) were released in 1966, 1972, and 1992. Agile practices bring novel challenges; some of them emphasize early and continuous delivery, which may contradict or hinder careful ethical thinking, planning, and execution (Spiekermann (2019), Gibson et al. (2022)). Certain agile practices, however, might be well-suited to identify

potential issues; for instance, having business representatives and end users work with the development team on a daily base reduces the risk of misunderstanding and failing to meet their expectations. Ethics are not mentioned explicitly but touched upon in the "Manifesto for Agile Software Development" from 2001, which is based on four value statements itself; technical excellence is established as one of twelve principles in the Manifesto. Working software is the primary measure of progress and success, not its ethical properties.[5]

Current state of research. Many researchers highlight the relevance of ethics in software engineering and the threats posed by recent developments in related fields such as artificial intelligence, big data and Web development. An IEEE Software editorial positioned ethics as a "software design concern" (Ozkaya (2019)). Hole (2019) called for five principles: "ensure openness, avoid lock-in, pay for user information, provide multiple solutions with similar services, and combine minds and machines." Safety and privacy as well as robustness have received more attention than other values so far, for instance in IEC 61508 and the General Data Protection Regulation (GDPR).[6] Application domains differ in their adoption and maturity w.r.t. these values and qualities; e.g., software controlling medical devices can be expected to do better than situational apps for leisure and entertainment.

The Software Engineering Body of Knowledge (SWEBOK)[7] picks up the ACM and IEEE codes. In many countries, ethics education receives increasing attention in computer science and software technology curricula (ACM and IEEE Computer Society (2013), Dodig-Crnkovic and Feldt (2009)). The gray literature also raises awareness. An example of valid but rather generic and abstract advice to practitioners is to focus on service delivery quality (of people) and "act with integrity" and value "respect, trust, responsibility" (Hall (2009)). The recently published standard IEEE 7000-2021, "Standard Model Process for Addressing Ethical Concerns during System Design", defines five analysis and design processes to support this advice; it also suggests (but does not norm) an initial value catalog (IEEE 2021).

Few research projects address the problem domain from a method engineering or design science point of view; managing ethical values and risks on agile projects has received little attention so far. Issues have been reported (Gregory and Taylor (2013), Dindler (2022)) and the connection between technical debt and ethics has been identified (Gibson et al. (2022)). Economics researchers define digital value systems (Spiekermann (2019), Diethelm and Sennhauser (2019)).

In summary, existing work has focused on creating awareness. It followed a passive, document-oriented approach requiring project teams to pull knowledge and advice from the literature; methods and tools to stimulate ethically responsible behavior are missing. We propose to overcome these deficits by integrating ethical values into contemporary agile development routines. We do so in the form of an extended set of agile practices. We contribute an active push approach that makes the elicitation and prioritization of ethical values mandatory, effectively bringing value-based design into development workflows.

Results. Our *Ethical Software Engineering (ESE)* balances both human values such as fairness and diversity with agile values such as customer collaboration and responding to change. We inject value-based ethical engineering in the agile software development mainstream by way of

[5] https://www.agilealliance.org/agile101/the-agile-manifesto/ (2001)

[6] https://en.wikipedia.org/wiki/IEC_61508 and https://eugdpr.org/

[7] https://www.computer.org/education/bodies-of-knowledge/software-engineering

a novel approach to method engineering and tool design. Our contributions fall in three categories:

1. *Knowledge.* A compilation of essential questions to ask on agile development projects, derived and distilled from existing software engineering codes of ethics and professionalism as well as related sources on value-based software engineering and agile coaching (Agile Alliance (2022), IEEE (2021)). This compilation is disseminated in the form of two novel agile practices called *Story Valuation* and *Ethics Review*; the existing practice of user storytelling is amended with value information complementing the business benefits in the "so that" part of the story template (that also has "As a [role]" and "I want to [capability]" parts).
2. *Methods.* We envision a decision support and tradeoff method for value-based resolution of conflicts between ethical and other design concerns. This method adopts and complements the process defined in IEEE 7000 (IEEE (2021)), working with its ConOps, value register, Ethical Value Requirement (EVR) and Value-Based System Requirements (VBSRs) artifacts. Existing agile practices for requirement prioritization, project planning, and reflection (e.g., definition of ready, definition of done, retrospective) are updated; we also integrate the existing agile concepts of product backlog, sprint planning, and acceptance testing. Each ethically desired behavior is distilled a) from the existing body of knowledge and b) current project context and requirements. Values and resulting requirements are articulated in several different formats that are inspired by the agile user story template, including value narratives, value weightings and decision-oriented "context-criteria-options" triples. Such template-based value statements help to raise awareness for ethical concerns and make it harder to behave unethically. To stimulate ethical thinking even further, we also envision concrete, actionable conflict resolution advice that leaves professional responsibility with the engineer (where it belongs) but moderates the decision-making process.
3. *Tools.* We experimented with a demonstrator of a continuous ethics linter as a first tool that actively places ethical awareness in the development mainstream. This tool looks for ethical smells (i.e., suspects that a value might be harmed), inspecting source code and supplemental artifacts in project repositories. A first, basic, text-based prototype of such a linter tool unveiled technical feasibility but also ethical concerns; further research is required to set an adequate direction here.

   We validated our method engineering results in action research so far, with case studies and surveys planned; the project results are available in a public git repository at https://github.com/ethical-se. In our future work, we consider including pre-defined value catalogs and assessments of their relevance w.r.t. project phases and architectural layers (presentation, business logic, data access and storage) into our approach. We also consider developing additional templates and notations, emphasizing usability, scalability, and conflict management in our method engineering.

KEYWORDS: Agile Software Development, Design Decisions, IEEE 7000, Moral Values, Normative Ethics, Requirements Engineering, User Stories, Value-Based Systems Design.

REFERENCES

ACM. 2018. “ACM Code of Ethics and Professional Conduct 2018.”

ACM, and IEEE Computer Society. 2013. Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science. New York, NY, USA: ACM.

Agile Alliance. 2022. “Code of Ethical Conduct for Agile Coaching, Version 2.0.” https://www.agilealliance.org/agilecoachingethics

Diethelm, C., and P. Sennhauser. 2019. “Digitale Ethik, HWZ Whitepaper.”

Dindler, Krogh, C. 2022. “Engagements and Articulations of Ethics in Design Practice.” *International Journal of Design* 1 (2): 47–54. https://doi.org/10.57698/v16i2.04

Dodig-Crnkovic, G., and R. Feldt. 2009. “Professional and Ethical Issues of Software Engineering Curriculum Applied in Swedish Academic Context.” In HAoSE 2009 First Workshop on Human Aspects of Software Engineering. Online Proceedings. http://www.es.mdh.se/publications/1616

Gibson, John Paul, Massamaesso Narouwa, Damian Gordon, Dympna O’Sullivan, Jonathan Turner, and Michael Collins. 2022. “Technical Debt is an Ethical Issue.” Proc. of ETHICOMP 2022.

Gregory, Peggy, and Katie Taylor. 2013. “Social and Communication Challenges for Agile Software Teams.”, *Proceedings of ETHICOMP 2013*, 186–191.

Hall, D. 2009. “The Ethical Software Engineer.” *IEEE Software* 26 (4): 9–10. https://doi.org/10.1109/MS.2009.106

Hole, K. J. 2019. “Dominating Software Systems: How to Overcome Online Information Asymmetry.” *IEEE Software* 36 (4): 81–87. https://doi.org/10.1109/MS.2019.2903075

IEEE. 2021. “IEEE 7000 Standard Model Process for Addressing Ethical Concerns During System Design.”

Ozkaya, I. 2019. “Ethics Is a Software Design Concern.” *IEEE Software* 36 (3): 4–8. https://doi.org/10.1109/MS.2019.2902592

Spiekermann, S. 2019. *Digitale Ethik: Ein Wertsystem Für Das 21. Jahrhundert*. Droemer Verlag.

Whittle, Jon. 2019. “Is Your Software Valueless?” *IEEE Software* 36 (3): 112–15. https://doi.org/10.1109/MS.2019.2897397

THE DEMOCRATIZATION OF OUTER SPACE: ON LAW, ETHICS, AND TECHNOLOGY

Eleonora Bassi, Ugo Pagallo

Politechnic of Turin (Italy), University of Turin (Italy)

eleonorbassi@gmail.com; ugo.pagallo@gmail.com

EXTENDED ABSTRACT

The research addresses the challenges brought forth by projects on mass space exploration developed by private companies as well as current investments on space tourism, space hotels, and other space human activities, e.g., scientific research in outer space missions over the next few years. Such projects and investments go hand-in-hand with the growth of the space economy and business revenue hinging on dramatic decreasing costs for space missions and spacecrafts (Lyall and Larsen 2018; Ziemblicki and Oralova 2021). In turn, the growth of space economy and current activism of policy makers depends on the exponential advancements of technology, from smart robots equipped with AI to increasingly autonomous and intelligent artificial systems (Bratu et al. 2021; Martin and Freeland 2021; Pagallo et al. 2023).

The scenario of multi-planetary human life entails fascinating problems of political philosophy, ethics, and legal theory on how to govern millions of people in space. The focus of the analysis in this paper is restricted to current efforts of EU lawmakers to address the challenges of AI systems. The case study of the European Space Agency (ESA) and the arbitration clauses of its contracts for the use of the Columbus Laboratory in the International Space Station aims to illustrate the limits of traditional approaches, and why principles and provisions of space law should be complemented with further fields of legal regulation, such as those of personal data protection and privacy, cybersecurity and machinery regulation, down to tortious liability and consumer law (Pagallo 2011; Pagallo 2013a; Bassi et al. 2019; Falco 2019). The assumption is threefold. First, the quest for the democratization of outer space casts further light on the democratic deficit of such institutions, as the European Union, vis-à-vis current trends on the privatization of outer space. Second, ethics and moral arguments play a critical role in filling the gaps and shortcomings of current legal regulations, both contributing to shaping legislation and interpreting valid law in the best possible light (Marsh 2006; Pagallo 2013b; Pagallo 2018; Jessen 2017; Rogerson 2022). Lastly, from a legal viewpoint, it seems fair to admit that most liability issues of outer space (Ernest 1991; Dennerley 2018; Larsen 2019) will progressively regard private parties and safeguards that private companies should guarantee to protect the rights of the next generation of space tourists, explorers, and even settlers (Freeland and Jakhu 2014; Scheutz and Arnold 2016; Lim 2020; Freeland and Ireland-Piper 2022; Martin and Freeland 2022).

In order to provide a hopefully fruitful view on the subject matter, we plan to divide the analysis into four parts. First of all, focus will be on a main driver of the next generation of space tourists and explorers, namely, the dramatic decreasing costs for space missions and spacecrafts. Then, the analysis dwells on the core of the privatization of outer space, i.e., the very appropriability of space resources, as the crux of many debates of today's public international law (Pekkanen 2019). On this basis, the further step is to investigate how EU law regulates the status of potential outer space tourists, or explorers, with the case study of the regulatory framework for

the Columbus Laboratory in the International Space Station (ISS). Finally, the drawbacks of this regulatory framework are under scrutiny vis-à-vis current efforts of lawmakers to tackle the normative challenges of AI in such fields as cybersecurity, machinery safety, consumer law, data protection, and more.

Drawing on tenets of space law, philosophy of technology, ethics, and technological regulation, the conclusion of the investigation stresses the relevance of the issue, i.e., the 'democratization of space' and why the speed of technological innovation together with human ingenuity will increasingly put this topic in the spotlight. In 2022, the Director of the new heavyweight aerospace contractor SpaceX, i.e., Benji Reed declared in a press briefing what they want: "We want to make life multi-planetary, and that means putting millions of people in space." Leaving aside the promise, or the menace of Space X's Director on "millions of people in space," it seems fair to admit that we should be ready to properly tackle the challenges of this next generation of humans that will leave Mother Earth.

KEYWORDS: Artificial Intelligence; Democracy; Human Rights; International Space Station; Robotics; Space Law.

REFERENCES

Bassi, E., Bloise, N., Dirutigliano, J. et al. (2019) The Design of GDPR-Abiding Drones Through Flight Operation Maps: A Win–Win Approach to Data Protection, Aerospace Engineering, and Risk Management, *Minds & Machines*, 29, 579–601.

Bratu I., Lodder A.R. and T. van der Linden (2021) Autonomous space objects and international space law: navigating the liability gap, *Indonesian Journal of International Law*, 18(3): 423-446.

Dennerley, J.A. (2018) State liability for space object collisions: the proper interpretation of 'fault' for the purposes of international space law, *European Journal of International Law*, 29(1): 281-301.

Ernest, V.C. (1991) Third Party Liability of the Private Space Industry: To Pay What No One Has Paid before, *Case W. Rsrv. L. Rev.*, 41, 503-541.

Falco, G. (2019) Cybersecurity Principles for Space Systems, *Journal of Aerospace Information Systems*, 16(2): 61-70.

Freeland, S. and R. Jakhu (2014) What's human rights got to do with outer space?: everything!. In R. Moro- Aguilar, P. J. Blount, & T. Masson-Zwaan (Eds.), *Proceedings of the International Institute of Space Law 2014* 366.

Freeland, S. and D. Ireland-Piper (2022) Space law, human rights and corporate accountability. *UCLA Journal of International Law and Foreign Affairs*, 26(1): 1-34.

Jessen, D. (2017) Modern Ethical Dilemmas Stemming from Private One-Way Colonisation of Outer Space, *Journal of Space Law*, 41(1): 117-132.

Larsen, P. (2019). Commercial Operator Liability in the New Space Era. *AJIL Unbound, 113*, 109-113.

Lim, J. (2020). Charting a human rights framework for outer space settlements. *71st International Astronautical Congress (IAC)—The CyberSpace Edition*. https://www.jusadastra.org/assets/files/IAC-20,E7,2,11,x60311(1).pdf

Lyall, F. and P.B. Larsen (2018) *Space Law: A Treatise*, London, Routledge.

Marsh, M. (2006) Ethical and medical dilemmas of space tourism, *Advances in Space Research*, 37(9): 1823-1827.

Martin, A.-S. and S. Freeland (2021) The Advent of Artificial Intelligence in Space Activities: New Legal Challenges, *Space Policy*, 55, 101408.

Martin, A.-S. and S. Freeland (2022) A Round Trip to the Stars?: Considerations for the Regulation of Space Tourism, *Air and Space Law*, 47, 261-284.

Pagallo, U. (2011) Killers, fridges, and slaves: a legal journey in robotics. *AI & Soc*, 26, 347–354.

Pagallo, U. (2013a) Robots in the cloud with privacy: A new threat to data protection? *Computer Law & Security Review*, 29(5): 501-508.

Pagallo, U. (2013b) *The Laws of Robots: Crimes, Contracts, and Torts*, Springer, Dordrecht.

Pagallo, U. (2018) Vital, Sophia, and Co.—The Quest for the Legal Personhood of Robots, *Information*, 9, 230.

Pagallo, U., Bassi, E. & Durante, M. (2023) The Normative Challenges of AI in Outer Space: Law, Ethics, and the Realignment of Terrestrial Standards. *Philos. Technol.* **36**, 23.

Pekkanen, S. M. (2019) Governing the new space race. *American Journal of International Law*, *113*, 92-97.

Rogerson, S. (2022) Ethical Digital Technology in Practice, CRC Press, Boca Raton.

Scheutz, M. and T. Arnold (2016) Are we ready for sex robots? *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 351-358.

Ziemblicki, B.; Oralova, Y. (2021) Private Entities in Outer Space Activities: Liability Regime Reconsidered, *Space Policy*, 56, 101427.

TECHNO-HEALTHISM: BEING PATIENTS-IN-WAITING UNDER THE DEVELOPMENT OF MEDICAL TECHNOLOGIES

Ryoko Asai

Ruhr University Bochum, Bochum (Germany); Meiji University, Tokyo (Japan); Uppsala University, Uppsala (Sweden)

ryoko.asai@rub.de

EXTENDED ABSTRACT

AI has been incorporated into medical practice and is able to detect the risk of various illnesses in the future, based on analysis of past medical data and people's current health status and lifestyle habits. In order to reduce the estimated risk of illness, or in other words, to control the risk of developing illness in the future, preventive medical treatment may be offered by doctors, or preventive measures may be taken spontaneously by the person who is informed of the risk. This trend can be seen in the current growing attention to self-care. Self-care is not only needed for people to be healthy (disease-free), but also for people to live independently coping with illnesses such as lifestyle-related illnesses, through daily medications and lifestyle control by the individual. In other words, self-care is considered effective in enabling people to live as independently as possible, regardless of whether they are ill or not, without becoming bedridden (Asai, 2022). Also, this can be explained by the recent rise in health consciousness.

Given these circumstances, the use of medical AI and digital health is likely to become more active in the future, both in medicine and care. While we are very positive about the contribution of these emerging technologies to human health, how aware are we of the political and ethical implications of these technologies behind their technological functions? This study is not a discussion of the technological functions of medical AI or digital health, but rather an examination of the ethical implications of these emerging health-related technologies for our 'life'. Also, this study is questioning what 'health' means under the development of medical technologies. There has been a great accumulation of research, mainly by medical sociologists, on the social aspects of health (Timmermans and Haas, 2008).

In particular, the distinction between the social and biological aspects of health, as articulated by Parsons, has had a major influence on these studies (Timmermans and Haas, 2008; Timmermans and Buchbinder, 2010). The distinction between illness and disease has also been explained, where disease is defined as the experience of feeling sick, and illness as an organic and pathological condition based on a more medical diagnosis. Doctors have played a role of legitimising our experience of disease as a medical condition, with diagnosis based on their professional knowledge. Particularly in modern medicine, the shift from a view of disease as part of the patient to a view of disease as an independent entity has made the role of the doctor's diagnosis more important (Rosenberg, 2007). Timmermans and Buchbinder, drawing on previous research, describe diagnosis as both a process of deliberate judgement and a pre-existing set of categories that initiate a series of experiences, identities, life strategies and subsequent medical practices (Timmermans and Buchbinder, 2010; Rosenberg, 2007; Jutel, 2009).

When medical AI is introduced to diagnosis by doctors and the vast amount of records and data relating to illnesses are utilised, it appears that more precise and valid diagnoses as well as effective medical practices will become possible. For example, the situation of people who suffer from experience of diseases that interfere with their daily life, but are not diagnosed as illnesses by doctors, could be improved (Atkins, 2010).

Moreover, the early detection or risk prediction of illnesses by these advanced technologies is considered very useful for people who do not have any obvious experience of disease, but who want to stay healthy and consult a doctor for an assessment of their health status. However, especially the risk prediction of illnesses can encourage, and sometimes force, people to take preventive medication and to adopt supposedly 'healthy' habits and lifestyles in order to avoid the risk of illnesses that are yet to be discovered.

In other words, while the advancement of cutting-edge medical technologies could provide us with the knowledge of how to be more physically healthy, being physically healthy could become an end in itself. Physical health has been replaced from being one of the key means of achieving personal goals and living a meaningful and fulfilling life to being healthy as a meaningful life goal in itself. Whilst it is very natural to make being healthy one of our goals, it does not mean that being healthy is perfectly equal to being happy. Our wellbeing can be broadly categorised into subjective wellbeing and psychological wellbeing (Helliwell, Layard, Sachs, De Neve, Aknin, & Wang, 2022). Although physical health obviously affects both, there are still many people who feel happy even when they are ill, those who are physically healthy but never feel happy, and those who consume substances harmful to their health (e.g., drinking alcohol or smoking) but still feel happy. That is, in addition to our physical health, which can be measured by medical technology, our psychological state, which is difficult to measure by machines, is also a very significant factor in our well-being. What does it mean for happiness and wellbeing if the goal is to maintain physical health, and sometimes, for the sake of health, to give up everything that gives us pleasure and enjoyment? Such questions may bring us back to the philosophical question of 'what is happiness', which has been discussed since ancient Greece.

In addition, when people are informed in advance by their doctors about foreseeable risks of illness, patients may feel that they need to take more responsibility for their health and illness. The change in perception from illness being something that anyone can get to illness being something that can be predicted and prevented may require all of us to take more responsibility for our health. If this latter perception permeates society, it is inevitable that even the health and social care systems will be changed based on this perception. Individual responsibility and burden for health may be increased in the future in order to reduce the societal burden on healthcare and social welfare.

This situation has already been described by Skrabanek as the rise of healthism and lifestylism (Skrabanek, 1994). According to him, the doctrine of lifestylism considers that "most diseases are caused by unhealthy behaviour". He then also pointed out the moral and ethical problems that hover there, as follows:

> “Although lifestylism has a strong moral flavour, its language is mathematical. Each 'risk factor' has a number, which quantifies the risk”.

This description corresponds exactly to the calculated risk of illness and to the foreseen risk of illness. As medical AI diagnostics become more prevalent and the technology more developed, the mathematically calculated risks will have greater significance in people's daily lives. When

such lifestylism becomes more prevalent in society, people's attitudes to health are linked to political developments, as Skrabanek described in Healthism (Skrabanek, 1994):

> "The pursuit of health is a symptom of unhealth. When this pursuit is no longer a personal yearning but part of state ideology, healthism for short, it becomes a symptom of political sickness"

In societies where advanced technology is involved in people's health, healthism can be further amplified by the technology. This study attempts to explain such situations as techno-healthism.

Furthermore, in a society where such techno-healthism is concerned, we can all become patients-in-waiting (Timmermans and Buchbinder, 2010). In other words, we can be listed as potential or preemptive patients when technology provides a name or risk of illness to anyone who is not yet ill, who has not yet been diagnosed with a disease, or who is worried about health. The concept of patients-in-waiting was proposed by Timmermans and Buchbinder in their research on newborn screening tests (Timmermans and Buchbinder, 2010). It is derived to describe the interaction between newborns in an ambiguous condition with no clear diagnosis of disease, their parents and the doctors trying to diagnose them medically, and the situation in which they are placed.

Developments in medical AI and digitalhealth can impose a high degree of uncertainty on people who are pronounced at risk of future illness, even if they are healthy at the time. They would then accept preventive medical interventions to get out of the ambiguous situation and accept regular monitoring and tracking of their health status by their doctors. This means that we, who live in a precarious situation between healthy and sick, situate ourselves as patients and try to avoid the indeterminate situation (uncertainty) created by advanced medical technologies.

This study aims to examine the ethical problems posed by health, as defined by highly developed medical technologies and their mathematical processing, from the perspective of information ethics. In particular, it attempts to develop the idea of healthism into technohercism by focusing on the situation where the penetration of emerging technologies into the medical and care domain influences people's health consciousness, the societal meaning of 'health' and its ethical implications.

KEYWORDS: Ethics, Lifestylism, Medical AI, Patients-in-waiting, Techno-Healthism.

REFERENCES

Asai, R. (2022). Care and Data: How can we use healthcare data ethically? Proceedings of ETHICOMP 2022, 607-610.

Atkins, C.G.K. (2010). My Imaginary Illness: A Journey into Uncertainty and Prejudice in Medical Diagnosis. ILR Press.

Charmaz, K. (1991). Good Days, Bad Days: The Self in Chronic Illness and Time. Brunswick, NJ:Rutgers University Press.

Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., & Wang, S. (Eds.). (2022). World Happiness Report 2022. New York: Sustainable Development Solutions Network.

Jutel, A. (2009). Sociology of Diagnosis: A Preliminary Review. Sociology of Health and Illness, 31, 278–299.

Rosenberg, C. E. (2007). Our Present Complaint: American Medicine, Then and Now. Baltimore, MD:Johns Hopkins University Press.

Skrabanek, P. (1994). Death of Humane Medicine: And the Rise of Coercive Healthism. The Social Affairs Units.

Timmermans, S. and Buchbinder, M. (2010). Patients-in-Waiting: Living between Sickness and Health in the Genomics Era, Journal of Health and Social Behavior, 51(4), 408 –423.

Timmermans, S. and Haas. S. (2008). Toward a Sociology of Disease. Sociology of Health and Illness, 30, 659–676.

DOXING ETHICS

Juhani Naskali, Minna Rantanen, Maria Rottenkolber, Kai K. Kimppa

University of Turku, Turku School of Economics (Finland)

juhani.naskali@utu.fi; mimaran@utu.fi; maria.m.rottenkolber@utu.fi; kai.kimppa@utu.fi

EXTENDED ABSTRACT

Doxing is a practice where a third party, i.e., one or several doxer(s), intentionally publishes personal information about another individual, the doxee or target, without consent on the Internet (Douglas, 2016; Eckert and Metzger-Riftkin, 2020). The information revealed may include victims' real names, home addresses, or telephone number, among others. Thus, doxing can be considered a user-led violation of privacy (Trottier, 2017). In public discourse, doxing frequently holds an exclusively negative connotation (Barry, 2021), and academic literature has referred to doxing as a form of "problematic speech on social media" (Fleischman and Rosenbloom, 2020). While doxing may be intuitively condemned as a form of online harassment, there may be cases in which it serves as an ethically justified means of resistance (Cheung, 2021).

Utilitarianism is an ethical theory that advocates for actions to be judged based on their consequences. It posits that the moral worth of an action lies in its ability to maximize well-being, and to promote the greatest overall happiness or utility to the greatest number of people, often referred to as the "greatest happiness principle".

This paper investigates the phenomenon of doxing to answer the question of how to ethically evaluate instances of doxing from a utilitarian perspective. Specifically, this paper aims to examine whether doxing can be categorically considered ethically problematic or whether a more nuanced understanding of doxing is needed.

A utilitarian analysis of doxing must first identify the different types of doxing actions and the relevant groups of people whom are influenced by the consequences of doxing. Then, an analysis is conducted on the utility of the consequences of the identified types of doxing actions to these groups. The analysis includes indirect consequences, such as ridicule or financial and physical harm to the target.

There are many different ways to categorize doxing (Anguita, 2021; Cheung, 2021; MacAllister, 2017; Snyder et al., 2017). For the purposes of this paper, a high-level categorization is the most fruitful. Douglas (2016) identifies three main forms of doxing: 1) deanonymizing doxing, which identifies a previously anonymous or pseudonymous person; 2) targeting doxing, which makes it easier to physically locate and contact the target, possibly increasing the risk of harassment or physical harm; 3) delegitimizing doxing, which undermines the credibility of the target.

Information that suggests that the victim has breached a social norm or committed an immoral act, for instance, can undermine the target's reputation. While targeting doxing provides the means to harass the target, delegitimizing doxing can provide a 'reason' for harassment. (Douglas, 2016.)

Special cases of doxing have been identified previously. For example, if an individual uses anonymity as a means to avoid being held accountable for wrongdoing or to mislead others, the

public might have a legitimate interest in revealing this wrongdoing by uncovering the person's identity and thereby removing anonymity (Douglas, 2020). This disclosure increases the public's ability to hold the wrongdoer accountable as such doxing can help stop such wrongdoing in the future and signal to the broader community that such wrongdoing is not tolerated and has negative consequences.

The affected groups in doxing are a) the target of doxing, b) people close to the target, such as friends and family, who can be directly affected by the situation, c) the doxer, d) people who benefit from the doxing (e.g., when doxing reveals illegal misdeeds), whether legitimately or illegitimately, and e) the public at large.

In deanonymizing doxing, the target suffers the loss of protection that anonymity provides (1a). This can lead to online harassment and threats, resulting in emotional distress, fear, a sense of insecurity and loss of privacy. Additionally, their personal and professional life may be adversely affected by the reactions of their friends, families and employers, leading to reputational damage or job loss. Douglas (2020) posits that any shaming that accompanies doxing is only permissible if it is reintegrative: *"Without the possibility that those who are exposed by digital vigilantism can be reintegrated into their communities, DV risks further alienating them and reinforcing their extreme views."* From a utilitarian perspective, permanent consequences have a much stronger utility (whether negative or positive). Sometimes harassment and threats can bleed over to the close ones of the target, who can also face harassment or threats due to their association with the target (1b). The doxer presumably has a goal in mind with their doxing, which yields utility to them (1c). If there are people who directly benefit from the information released with the doxing (e.g. financial records proving people were subject to fraud would benefit the victims, as well as the investigators), they may receive positive utility from it. The public at large may benefit from doxing that combats negative behavior (1d) if the behavior stops, and/or people are less likely to repeat the behavior in the future due to it being costly.

Targeting doxing holds the same negative utility for the target of the doxing (2a), and in addition, holds the possibility of physical harm, stalking and other offline harassment, while simultaneously holding a greater chance of psychological trauma, fear and diminished quality of life. The same applies to their close ones (2b). Unless their goal is physical harm, utility to the doxer does not change (2c). In cases where the target poses a genuine threat to others (e.g. terrorism), targeting doxing can have a high positive utility to law enforcement and potential victims of the target (2d). Utility to the public is likely to suffer with the inclusion of location information, as the threat of physical harm can lead to a culture that legitimizes vigilantism, harassment and the fear they lead to. In summary, targeting doxing has a much lower net utility than deanonymizing doxing unless it leads to the prevention of great tragedies, such as in the case of preventing physical calamities such as school shootings or terrorist attacks.

Delegitimizing doxing holds a larger negative utility for the target, in comparison to deanonymizing doxing, as they further take a hit to their credibility (3a). Delegitimizing doxing is an "attempt to shame and humiliate the subject, often by portraying her as a transgressor of an established (or supposed) social norm" (Douglas, 2016), and as such, it can weigh heavily on the target, and the change of social ostracism and loss of financial opportunity is much greater. Their close ones are similarly affected negatively (3b). While the doxer might experience a stronger sense of schadenfreude compared to deanonymizing doxing, it is generally short-lived and superficial (3c). It is difficult to imagine circumstances where a group of people would be positively affected by delegitimizing doxing—surely such utility is provided by evidence or

wrongdoing or the acts of the doxing target, and not from the delegitimizing (at worst, dehumanizing) the target. Even deanonymizing doxing can lead to delegitimizing the target, but delegitimizing doxing is done for the express "intention of undermining the target's credibility, reputation and/or character" (Douglas, 2016), which brings no further utility to others (3d). Delegitimizing doxing can even harm the public in "maintaining the 'tyranny of majority' that concerned John Stuart Mill" (Douglas, 2016), contributing to a culture of public shaming (3e). While some members of the public can view this as a means to expose hypocrisy, such acts can have negative consequences on free expression and public discourse that ultimately leads to better understanding.

Interestingly, the group with the most variability in the utility of doxing is d) the people who possibly benefit from the doxing. If doxing prevents a school shooting, for example, the negative utility of multiple shooting victims greatly outweighs the negative utility of the doxing target's loss of reputation and likely incarceration, were such a situation prevented. Outside of such extreme examples, doxing does not yield any positive consequences to the public, outside of deterrence to breaking social norms (general utility to e), the public), and it becomes much harder to overshadow the negative consequences to the target, especially in the case of targeted doxing.

In accordance with Douglas (2016), the ethical analysis finds that morally permissible cases of doxing need to serve the public interest without violating several side-constraints. The benefits to the public of exposing the victim's wrongdoing need to outweigh the harms to the victim. Douglas' original conceptual analysis was criticized by Barry (2021) to the extent that Douglas's specific consequentialist approach remains unclear, and that it is not explicitly stated what makes a public interest "compelling". We hope that this more detailed utilitarian analysis helps alleviate these concerns.

To conclude, balancing the potential benefits of accountability, deterrence, and public safety with the potential harms of privacy invasion, reputational damage, and emotional distress is crucial to addressing doxing in a responsible and balanced manner. For doxing to be morally permissible from a utilitarian viewpoint, it needs to bring benefit to others—either in reconciling a previous wrong or preventing future suffering.

KEYWORDS: Doxing, social media, freedom of speech, utilitarianism, ethics.

REFERENCES

Anguita, P. (2021). Freedom of expression in social networks and doxing. In *The Handbook of Communication Rights, Law, and Ethics: Seeking Universality, Equality, Freedom and Dignity*, pages 279-291. John Wiley & Sons.

Barry, P. B. (2021). Doxing racists. *The Journal of Value Inquiry*, 55(3):457- 474.

Cheung, A. (2021). Doxing and the challenge to legal regulation: When personal data become a weapon. In Bailey, J., Flynn, A., and Henry, N., editors, *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*, pages 577-594. Emerald Publishing Limited.

Douglas, D. M. (2016). Doxing: a conceptual analysis. *Ethics and Information Technology*, 18(3):199-210.

Douglas, D. M. (2020). Doxing as audience vigilantism against hate speech. In Trottier, D., Gabdulhakov, R., and Huang, Q., editors, *Introducing Vigilant Audiences*, volume 259, pages 259-280. Open Book Publishers Cambridge.

Eckert, S. and Metzger-Riftkin, J. (2020). Doxxing. In Ross, K., Bachmann, I., Cardo, V., Moorti, S., and Scarcelli, M., editors, *The International Encyclopedia of Gender, Media, and Communication*, pages 1-5. Major Reference Works.

Fleischman, W. and Rosenbloom, L. (2020). Problems with problematic speech on social media. In Pelegr´ım-Borondo, J., Arias-Oliva, M., Murata, K., and Palma, A. M. L., editors, *Paradigm Shifts in ICT Ethics: Proceedings of the Ethicomp 2020*, pages 116-120.

MacAllister, J. M. (2017). The doxing dilemma: seeking a remedy for the malicious publication of personal information. *Fordham Law Review*, 85(4):2451- 2483.

Snyder, P., Doerfler, P., Kanich, C., and McCoy, D. (2017). Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *Proceedings of the 2017 Internet Measurement Conference*, IMC ’17, pages 432-444. Association for Computing Machinery. event-place: London, United Kingdom.

Trottier, D. (2017). Digital vigilantism as weaponisation of visibility. *Philosophy & Technology*, 30(1):55-72.

A PRELIMINARY SURVEY OF MANUFACTURING WORKERS ABOUT AI IN THEIR WORKPLACE

Andrew A. Adams, Iraide Unanue Calvo

Meiji University (Japan), Tecnalia (Spain)

aaa@meiji.ac.jp; iraide.unanue@tecnalia.com

EXTENDED ABSTRACT

AI, after more than fifty years of promising "major useful breakthroughs within a decade or two" (Simon, 1960) finally delivered on that promise in the mid-2010s with increasing academic and news coverage. There has been both a wide development of applications, and an increasing level of attention of its impact on society, positive and negative, beyond the existing body of literature on speculative ethics analyses (AI & Society, 1987; Amigoni, Schiaffonati & Somalvico, 1999; Floridi, 2008). In addition to the direct social issues it produces, the economic impact of a wave of additional automation in workforces has also been considered speculatively for decades. Very recently the breakthrough in generative AI (Haleem, Javaid & Singh, 2022), has received considerable attention about the potential impact on employment in areas such as copywriting (Zarifhonarvar, 2023) and journalism (Pavlik, 2023). The potential impact on manufacturing has been the subject of economic discussion since the 1990s (Rifkin, 1996), but limited empirical study, and almost none focussing on workers, although the OECD (Lane, Williams & Broecke, 2023) recently surveryed both employers and workers. In the hope of inspiring more and better research in this area, a survey was undertaken across eight countries, with 1082 non-managerial participants from manufacturing.

The survey was deployed in summer 2022 in three German-speaking countries (Austria, Germany and Switzerland), two other large European countries (Spain and the UK), two smaller European countries (Slovenia and Greece) and in Japan. The choice of countries was largely driven by practical issues. The survey was developed in English then translated into German, Greek, Japanese, Slovenian and Spanish by native-speaking researchers. Survey participants were recruited using professional recruitment firms. All countries were surveyed using Internet surveys except for Slovenia in which the survey was administered by telephone. A reasonable gender balance of participants was sought, only Japan had a 50/50. The most gender unbalanced was in the German-speaking countries with only 28% of respondents being female. (Non-binary gender was offered as an option in the survey but no respondents selected this option.) See Table 1 for breakdowns of gender and participants per country.

In addition to demographic data (including work history and expectations) 19 questions about knowledge of/attitudes towards AI for Manufacturing and its social/economic implications with a seven point Likert answer scale from "very strongly disagree" to "very strongly agree". As an exploratory survey these questions were presented to the respondents without explanation or definition of AI.

Table 1. Participants by Country and Gender.

Country/ies	**Number**	**Male % (N)**	**Female % (N)**
German-speaking (DE, OS, CH)	200	72% (145)	28% (55)
Greece	109	64% (70)	36% (39)
Japan	222	50% (110)	50% (112)
Slovenia	307	60% (183)	40% (124)
Spain	121	64% (77)	36% (44)
UK	123	63% (78)	37% (45)
Total	**1082**	**61% (663)**	**39% (419)**

Summaries of total and per-country analysis of the responses to the attitude questions are presented below.

- AI technology is already important in my workplace.
  Moderate agreement overall with high agreement in Greece and German-speaking countries, moderate agreement in Spain and the UK, and moderate disagreement in Slovenia and Japan.
- AI technology will become important, or increase in importance, in my workplace over the next four years.
  Moderate agreement overall with high agreement in Greece, German-speaking countries, Spain and the UK, moderate agreement in Slovenia but neutral in Japan.
- I understand how AI technology can be used in the kind of work I do.
  Moderate agreement overall with high agreement in Greece, German-speaking countries, and the UK, moderate agreement in Spain and Slovenia, and neutral in Japan.
- I have experience with the use of AI related to my job.
  Neutral overall, with high agreement in Greece and German-speaking countries, neutral in Spain and the UK, and moderate disagreement in Slovenia and Japan.
- AI technology can improve the quality of the work I do.
  Moderate agreement overall with high agreement in Greece, German-speaking countries, Spain and the UK, moderate agreement in Slovenia, and neutral in Japan.
- AI technology could help me to become a more productive worker.
  Moderate agreement overall with high agreement in Greece, German-speaking countries, and Spain, moderate agreement in the UK, and neutral in Slovenia and Japan.
- AI technology could help my workplace become more inclusive.
  Moderate agreement overall with high agreement in Greece and German-speaking countries, moderate agreement in Spain and the UK, and neutral in Slovenia and Japan.
- AI technology could replace a large part or all of the job I currently do.
  Neutral overall, with high agreement in German-speaking countries and Greece, neutral in Spain and the UK, and moderate disagreement in Japan and Slovenia.
- The introduction of AI technology will help me to keep my job.
  Neutral overall with high agreement in Greece and German-speaking countries, neutral in Spain, the UK and Japan, and moderate disagreement in Slovenia.

- If AI technology replaces part or all of my current job, my employer will retrain me, and anyone else doing similar jobs, into other work with similar pay and conditions.
  Neutral overall with high agreement in Greece and German-speaking countries, neutral in the UK, Japan, and Slovenia, and moderate disagreement in Spain.
- If AI technology replaces part or all of my current job, my employer will make some or all of the people doing these jobs redundant.
  Moderate agreement overall with high agreement in Greece, moderate agreement in German-speaking countries, the UK, Slovenia and Spain, and neutral in Japan.
- I will find it easy to get a replacement job with similar pay and conditions if I am made redundant.
  Moderate agreement overall with high agreement in Greece and German-speaking countries, moderate agreement in Spain, neutral in the UK and Slovenia and moderate disagreement in Japan.
- AI will create more jobs than it will eliminate.
  Neutral overall with high agreement in Greece, moderate agreement in German-speaking countries, neutral in Japan and Spain and moderate disagreement in the UK and Slovenia.
- I would move to a new job in order to work with up-to-date manufacturing technology.
  Moderate agreement overall with high agreement in Greece and German-speaking countries, moderate agreement in Spain, Slovenia and the UK, and moderate disagreement in Japan.
- My employer has provided me with adequate training to use new manufacturing technology.
  Moderate agreement overall with high agreement in Greece and German-speaking countries, moderate agreement in Spain and the UK, neutral in Slovenia, and moderate disagreement in Japan.
- It is important that humans have final control when AI technology is used in manufacturing.
  High agreement overall with high agreement in all countries.
- Companies should pay equivalent taxes for AI/robotic workers if they reduce their human workforce.
  Moderate agreement overall with high agreement in Greece, German-speaking countries, Slovenia and Spain, moderate agreement in the UK and neutral in Japan.
- I understand the idea of a Universal Basic Income system.
  Moderate agreement overall with high agreement in German-speaking countries, Greece and Spain, moderate agreement in Slovenia and the UK and moderate disagreement in Japan.
- I support the introduction of a Universal Basic Income (UBI) or similar system.
  Moderate agreement overall with high agreement in Greece and German-speaking countries, moderate agreement in Spain, Slovenia and the UK, and neutral in Japan.

KEYWORDS: Artificial Intelligence; Manufacturing; Workers; Workforce Impact.

ACKNOWLEDGEMENTS: This survey was carried out as part of the EU-Japan.AI project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957339. Thanks to Kiyoshi Murata and Yasunori Fukuta of Meiji University, Gisela Hagmair and Damir Haskovic of Minds and Sparks, Matej Kovacic of JSI, and George Siaterlis of LMS, for translations and other support in deploying and analysing the survey.

REFERENCES

AI & Society (1987). Editorial. AI & Society, 1 pp. 3-4. https://doi.org/10.1007/BF01905884

Amigoni, F., Schiaffonati, V., & Somalvico, M. (1999). Some ethical aspects of agency machines based on artificial intelligence. In Proceedings of the Fourth ETHICOMP International Conference on Social and Ethical Impacts of Information and Communication Technologies (pp. 6-8). Retrieved from https://schiaffonati.faculty.polimi.it/pubblicazioni/C2.pdf

Floridi, L. (2008). Artificial intelligence's new frontier: Artificial companions and the fourth revolution. Metaphilosophy, 39(4-5), 651-655.

Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. BenchCouncil transactions on benchmarks, standards and evaluations, 2(4), 100089. https://doi.org/10.1016/j.tbench.2023.100089

Lane, M., Williams, M. & Broecke, S. (2023), "The impact of AI on the workplace: Main findings from the OECD AI surveys of employers and workers", OECD Social, Employment and Migration Working Papers, No. 288, OECD Publishing, Paris. https://doi.org/10.1787/ea0a0fe1-en.

Parsable (2021) The State of the Connected Frontline Manufacturing Worker, 2021. Retrieved from https://parsable.com/wp-content/uploads/2021/11/Global-FrontlineWorkerSurveypdf.pdf

Pavlik, J. V. (2023). Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. Journalism & Mass Communication Educator, 10776958221149577.

Rifkin, J. (1996). The End of Work. North Hollywood, CA, USA: Pacifica Radio Archives. Retrieved from http://pinguet.free.fr/rifkin1995.pdf

Simon, H. A. (1960). The new science of management decision. Harper & Brothers. https://doi.org/10.1037/13978-000

Zarifhonarvar, A. (2023). Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. Available at SSRN 4350925.

(REF)USING AI

Dr. Katleen Gabriels. Prof. Dr. Karl Verstrynge, Judith Campagne

Maastricht University (Netherlands), Vrije Universiteit Brusse (Belgium), Vrije Universiteit Brusse (Belgium)

k.gabriels@maastrichtuniversity.nl; karl.verstrynge@vub.be; judith.van.lookeren.campagne@vub.be

EXTENDED ABSTRACT

Introduction

AI is becoming an increasing part of our daily lives, one reason being that it is included in more and more 'smart technologies'. Policy decisions about city-management, for example, are turning progressively 'smarter' and big data driven: from smart waste management to smart parking. Frequently, the justification behind using big data is that it leads to more effectivity and an improved quality of life, such as better water pressure in city households (Kirstein et al., 2021).

However, the inclusion of smart technology also causes privacy concerns (e.g., Mihaljevic et al., 2021; Richards, 2013; Roessler, 2015; Zuboff, 2019). Being surrounded with cameras and sensors in the digital age, how critically can one still engage with these technologies? How much choice does one have when deciding to participate? For a while, discourses about digital inclusion were about being or having to *become* a user. The thought behind this was that digital exclusion ought to be avoided. The only subject-positions were that of participant and soon-to-be participant. Scholars such as Sally Wyatt, Anne Kaun, and Emiliano Treré are researching what it means for people to *not* partake in the digital society (Wyatt, 2003; Kaun & Treré, 2020). This shows that there is not a single way of engaging with technologies. Yet, in these perspectives, a binary way of thinking is sometimes still in place. To Wyatt (2003), again just two subject-positions seem to be available: that of user or non-user.

A binary understanding of use/non-use is already being challenged. To Finn Brunton and Helen Nissenbaum, for instance, the term 'obfuscation' helps to better understand how people are navigating the options between using and refusing technologies (2011). This is important, not the least, to highlight more ground for critical engagement with technologies, in the case of this presentation, more critical engagement with AI. To Marcuse (1969), for example, liberation from domination (for instance from oppressive forms of surveillance), requires thinking of new alternatives to the current ways society is being organized and how people move through it. To Marcuse (1969), this entails "a break with the familiar, the routine ways of seeing, hearing, feeling, understanding things so that the organism may become receptive to the potential forms of a nonaggressive, nonexploitative world" (p.6). Marcuse (1969, p.19) believes technologies play a crucial role in reshaping society in such a way that it moves away from exploitation and domination by materializing values such as freedom. This presentation takes that insight as a starting point and builds on current literature that conceives of various ways in which people move between using and refusing technologies. The aim of our approach is to give more space to the various options of engagement that reside between using and refusing AI. Such an overview will help in conceiving of further, liberatory ways of engaging with AI.

Whilst this presentation is conceptual in nature, we refer to specific examples too. One that shows how the lines between user and non-user can be blurred is clothing brand Cap_able (2023). This brand wants to make consumers aware of privacy, both as a moral value and a human right. The clothes include technologies that, when scanned by a smart technology camera on the street, show a picture of an animal instead of the clothing wearer's face. From the perspective of face recognition cameras, Cap_able's clothes are a way of walking through a public place in privacy. In doing so, Cap_able refuses some consequences, such as a loss of privacy, that come with smart technologies that are all around us. This example shows how people can use technologies to refuse participating in others. At the same time, this example raises critical questions. After all, not everyone can afford these clothes. What does it mean to think of a society in which everyone can choose to be unseen by face recognition cameras?

The aim of this presentation is to complement and add to recent conversations regarding the critical, public engagement with AI. An overview of some ways in which people are engaging with AI in a manner that moves between using and refusing the technology is a good starting point to think of the plurality of ways in which one can critically interact with AI. This is important, since living in a smart city can also mean resisting the smart city, or rather, can also mean resisting the digitalization's monopoly on what 'smart' means in the context of city-design. To have a healthy, digital society, people must have the opportunity to resist 'being smart' too. Yet, who has access to the technologies and practices of refusing these? Who can afford to buy specific clothing to resist face recognition cameras? It is precisely these types of questions that come to the fore when looking at a broader spectrum of critical use of AI.

State of the art

Until now, reflections on refusing technology are often framed around the perspective of non-use. ''Analyzing users is important, but by focusing on users and producers we run the risk of accepting a worldview in which adoption of new technology is the norm'' (Wyatt, 2003, pp.77-78). To Wyatt (2003), non-users are resisters, rejecters, the excluded, and the expelled. Resisters and the excluded are those who do not make use of a specific technology at all. The former because they do not want to, the latter because they cannot. Rejecters once used a technology but are now not keen on doing so anymore. Resisters have never used a specific technology; rejecters once did so but now (voluntarily) not any longer (Wyatt, 2003, p.76). The expelled once used technology but not anymore, because of involuntary reasons.

To Verdegem & Verhoest (2009), the list of non-users should not be exhaustive. They stress the importance of not viewing non-users as a homogenous group (Verdegem & Verhoest, 2009, p.650). Non-use is often seen as a tool towards resistance (Saxena et al., 2020). In those instances, it is indeed good to not consider the non-users to be a homogenous group. Yet, what happens when complete non-use is not possible? In those cases, one might use practices of obfuscation. Obfuscating software means using the technologies' methods to create within it a self-defeating process. In this manner, one uses the technology to resist the technology, for example by providing so much data that the system cannot possibly process it all (Brunton & Nissenbaum, 2015, p.18). Therefore, we claim that rejection, refusal, and obfuscation are ways of critically *using* technologies too.

More forms of critically (ref)using technologies

In addition to refusing, rejecting, resisting, and obfuscating, users also have the power to 'fit' or 'tweak' the technology, for instance by using the technology in deviating ways than the script the developers intended. Kamphof (2017) observed how caregivers who used monitoring technologies to observe their patients sometimes deliberately ignored data presented by the monitor if it was not in line with what the patient told them, so to respect the patient's right to their own version of a story. Privacy was reconsidered through a process of looking at a specific context and the role technologies play in that. Users might use a specific technology without adopting its original script.

Another form of showing resistance to surveillance technologies is by uniting consumers and starting a negotiation process together. A recent example is a group of Dutch schools who grouped together and successfully negotiated with big tech-companies such as Google (Alphabet Inc.) and Zoom to obtain better privacy conditions (Singer, 2023). Negotiation processes can turn the user into a non-user of the specific terms of the companies creating those technologies, whilst thereby actively establishing one's own terms.

In practices of cheating and protesting, users do not enter into a conversation with the designers of their technologies. Cheating is the act of deliberately confusing the data collected by a smart technology, such as giving the activity tracker to one's dog or using a device to stimulate movement on a laptop keypad, so that to an employer, it seems one is constantly working. Protesting is an expression of disapproval. This expression is often given form by calling on politicians to forbid the presence of a specific technology, or by activists to take matters in their own hands, for example by designing a system that can block Google Glass wearers from WiFi-networks (Newman, 2014).

The above shows various forms of critical engagement with smart technologies. In our final presentation, we will present an in-depth overview of the key terms refusing, resisting, rejecting, obfuscation, fitting, tweaking, cheating, negotiating, and protesting in relation to AI. Ultimately, we seek to show that a nuanced, conceptual dissecting establishes a thorough understanding of what the practices through which people are already examining these questions in their day to day lives look like. Creating such an overview of concepts related to critically engaging with AI opens up more space for and encourages an imaginary that can in turn again conceive of further practices of critically engaging with AI.

KEYWORDS: AI, Refusal, Resistance.

REFERENCES

Brunton, F. & Nissenbaum, H. (2011). "Vernacular resistance to data collection and analysis: A political theory of obfuscation." *First Monday 16*(5). https://doi.org/10.5210/fm.v16i5.3493

Brunton, F. & Nissenbaum, H. (2015). *Obfuscation: A User's Guide for Privacy and Protest*. MIT Press.

Cap_able. (n.d.). *Our mission*. Capable.design. https://www.capable.design/mission

Kaun, A. & Treré, E. (2020). "Repression, resistance and lifestyle: charting (dis)connection and activism in times of accelerated capitalism." *Social Movement Studies 19*(5-6): 697-715. https://doi.org/10.1080/14742837.2018.1555752

Kirstein, J.K., Høgh, K., Rygaard, M., Borup, M. (2021). "A case study on the effects of smart meter sampling intervals and gap-filling approaches on water distribution network simulations." *Journal of Hydroinformatics, 23*(1): 66-75.

Marcuse, H. (1969). *An Essay on Liberation.* Bacon Press.

Mihaljevic, H., Larsen, C.J., Meier, S., Nekoto, W., Zirfas, F.M. (2021). "Privacy-centred data-driven innovation in the smart city. Exemplary use case of traffic counting." *Urban, Planning and Transport Research, 9*(1): 425-448.

Newman, J. (June 4, 2014). 'Glasshole' Detector Blocks Google Glass Users' Wi-Fi. *TIME*. https://time.com/2822063/google-glass-glasshole-wifi-block/

Richards, N.M. (2013). "The Dangers of Surveillance". *Harvard Law Review 126*: 1934-65.

Roessler, B. (2015). Should personal data be a tradable good? On the moral limits of markets in privacy. In B. Roessler & D. Mokrosinska (eds.). *Social Dimensions of Privacy: Interdisciplinary Perspectives.* Cambridge University Press, pp. 141-161.

Saxena, D., Skeba, P., Guha, S. & Baumer, E.P.S. (2020). "Methods for Generating Typologies of Non/use." *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW1), Article no. 27. https://doi.org/10.1145/3392832

Sheehan, K.B. (2002). "Toward a Typology of Internet Users and Online Privacy Concerns." *The Information Society, 18*(1): 21-32. https://doi.org/10.1080/01972240252818207

Singer, N. (January 18, 2023). How The Netherlands is Taming Big Tech*. The New York Times*. https://www.nytimes.com/2023/01/18/technology/dutch-school-privacy-google-microsoft-zoom.html.

Verdegem, P. & Verhoest, P. (2009). "Profiling the non-user: Rethinking policy initiatives simulating ICT acceptance." *Telecommunications Policy 33*: 642-652. https://doi.org/10.1016/j.telpol.2009.08.009

Wyatt, S. (2003). Non-Users Also Matter: The Construction of Users and Non-Users of the Internet. In N. Oudshoorn & T. Pinch (eds.). *How Users Matter: The Co-construction of Users and Technology*. MIT Press, pp. 67-79.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Public Affairs.

ON THE CURRENT STATUS AND ISSUES OF PROGRAMMATIC ADVERTISING: PROSPECTS FOR MARKETING ETHICS

Hiroshi Koga

Kansai University (Japan)

hiroshi@kansai-u.ac.jp

EXTENDED ABSTRACT

1. Introduction

The purpose of this paper is to summarize the challenges of “programmatic advertising” and examine possible solutions. To that end, this paper is organized as follows. First, programmatic advertising is explained. Next, we summarize the issues of operational advertising. Next, solutions to the issues in Japan are presented. Finally, we present the challenges and future prospects for solving the problems.

2. The concept of programmatic advertisement

The concept of managed advertising generally refers to a method of placing advertisements in the most appropriate ad spaces based on the relevance of the budget, site, and ad content, without specifically fixing ad spaces. Busch (2014, p.8) lists the following five characteristics These are (1) granularity, (2) real-time transactions, (3) real-time information, (4) real-time creation, and (5) automation.

Programmatic advertising has been strong in recent years [Hackley and Hackley, 2019]. It refers to a type of advertising that delivers ads individually tailored to the interests of users based on the keywords they use when searching the Internet and the content of the websites they visit. Many people may have experienced relevant ads or pop-up ads on their social networking screens when surfing the Internet about beauty or searching for restaurants.

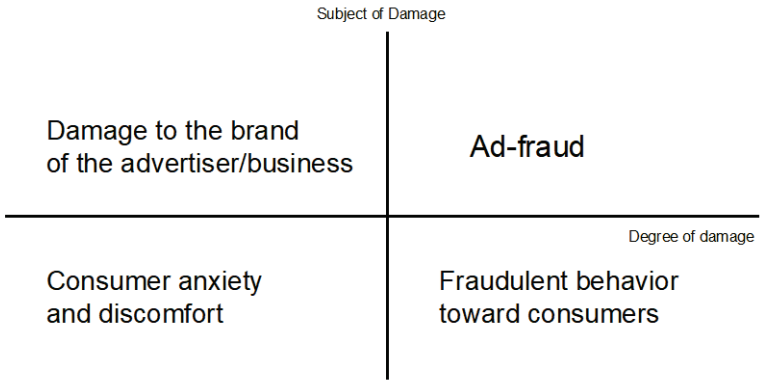
Thus, the advantages of managed advertising are that it is inexpensive, allows advertisements to be displayed to the appropriate target audience in real time, and allows the effectiveness to be monitored numerically. These characteristics are very different from those of traditional mass media advertising [cf. Palos Sanchez and Martin-Velicia, 2019].

3. The problems of programmatic advertisment

However, digital advertising is not a cure-all. A silver bullet is a weapon that can kill an immortal monster, the werewolf, with a single blow, and is made by melting down a cross [Markus and Benjamin, 1997]. From there, it is used to mean a panacea that solves difficult business problems in an instant. The authors would like to argue that programmatic advertising is not a cure-all. This is because there are several problems with programmatic advertising.

The author would like to argue that these problems can be classified into four major categories based on two dichotomies: the subject of advertising damage and the degree of damage (Figure 1).

Figure 1. Types of Problems with Operational Advertising.



First, brand damage to advertisers and companies refers to the risk of distributing advertisements to destinations that damage the image of the product, service, or company that is the subject of the advertisement. It is said to be caused by the development of ad technology, the monitoring system of platforms that automatically distribute ads in ad spaces, and the existence of fraudulent companies that establish unauthorized media.

The process of ad placement involves a large number of complex related players, called ad networks. It is automated. This means that the process is black boxed, making it difficult to identify and prevent the causes of fraud.

Next, ad-fraud is the practice of using automated programs to fabricate impression and click counts to fraudulently exploit the cost of performance-based advertising. It is sometimes referred to as “skip ads” or “backdoor ads”. In Japan, it is estimated that 3.3% of total advertising expenditures are funded by advertising (IAS, 2023). This means that 3% of advertising expenditures are wasted. For this reason, measures against ad-fraud are said to be an important issue.

Third, consumer anxiety and discomfort include the display of offensive ads, stealth marketing (ads posing as third parties), and ads that use “digital nudges” to induce purchases using psychological manipulation [Weinman, Schneider & Block, 2016].

Finally, there is fraud against consumers. Non-accredited institutions (degreed mills or diploma mills) that issue diplomas for money without providing proper education have become a social problem in the United States. This is said to be deeply rooted in advertisers who have developed methods (algorithms) to extract people with education complexes based on website browsing history and search terms. O’neil [2016] likens such harmful algorithms to weapons of mass destruction, calling them “weapons of mathematical destruction”.

By the way, it is not the responsibility of individuals to be fooled by weapons of mass destruction. Search sites such as Google tend to display relevant information based on our online activity history. As a result, we are comfortably ensconced in a bubble that uses our search history as a filter [Pariser, 2011]. In other words, we are ensconced in a comfortable bubble that displays only similar information and blocks out information we do not want to see (filter bubble). We are happy when we are exposed to information that says, “This is what I wanted to say”. However, when we are exposed only to certain information, our opinions are reinforced and we

become convinced that it is the only truth or righteousness, making us reluctant to accept opposing opinions (echo chamber phenomenon). This makes people more inclined to believe even fake advertisements (i.e., advertisements that use images of celebrities without their permission to create false testimonials).

4. Future prospects

Digital advertising in the filter bubble tends to create and reinforce desires as well as amplify people's opinions. It is not surprising that while searching for and browsing related articles about diet, people may become interested in frequent advertisements about a particular supplement. Subsequently, they may search for testimonials about supplements without knowing that they are ads, or they may search for native ads (ads created to look like online articles, which should clearly indicate that they are ads. It must be clearly marked as an advertisement). Thus, the pitfalls of managed advertising are not mutually exclusive, but interrelated.

However, as mentioned earlier, the companies involved in ad networks are diverse and intricately related, making it difficult to pinpoint the cause of problems when they do occur. In Japan, attempts are underway to avoid risks by creating a white list of groups of companies participating in ad networks. At the same time, there are moves to limit the damage to users by revising laws and guidelines.

However, the author believes that it is not easy to achieve marketing ethics through such institutional design alone. And as one practical solution, I would like to advocate the concept of "value creation through customer journey".

The "customer journey" is a visualization of the process by which a consumer is exposed to a certain product, becomes interested in that product through viewing various advertisements and online articles, and then makes a purchase. The customer journey is a graphical representation of the behavior and psychology leading up to the purchase of a product, assuming a specific consumer image (persona), such as a single part-time male in his 20s. To achieve this objective, digital advertising is noted to integrate the traditional distinction between ALT (mass advertising to increase product awareness) and BLT (non-mass advertising to stimulate purchase).

On the other hand, the customer journey method can be understood as forming a new use value for the product ex post facto through advertising, since the task is to strengthen the product image and stimulate and reinforce the desire to purchase through digital advertising. This is precisely the "competitive use value" pointed out by Ishihara (1982). Forgive me for using an old example, but Lotte once held a dance contest as a way of advertising its Fit Gum. The company succeeded in creating a product image and advertising by customers through the dance contest, not the gum itself. In this case, dance and music are new use values created by the advertising activity. This concept is competitive use value.

Malicious digital advertising, such as weapons of mass destruction, uses the temptation that the status quo will change at once if the product in question is purchased to convince customers that the consumption in question is a "cure-all" for them. Competitive use-value, on the other hand, becomes "co-creative use-value," in which value is proposed to customers and their reactions are monitored; in other words, value is created in collaboration with customers.

Also, whereas weapons of mass destruction stir up desire in a filter bubble, competitive (co-creative) use value creation has the potential to break through the filter bubble by creating new value. In other words, the mission of digital advertising in the future will be the co-creation of new values that will cause the scales to fall from our eyes, and the formation of communities that create a circle of empathy for these new values. In other words, digital advertising is required to create a value space coloured by a new sense of morality, and through this create "Alternatives to DX" that transforms the customer's life experience.

KEYWORDS: Programmatic advertising, marketing ethics, customer journey, co-creation value.

REFERENCES

Busch, O. ed. (2016). Programmatic advertising. *The Successful Transformation to Automated, Data-Driven Marketing in Real-Time.* Springer.

Hackley, C., & Hackley, A. R. (2019). Advertising at the threshold: Paratextual promotion in the era of media convergence. *Marketing Theory*, *19*(2), 195-215.

Isihara, T. (1982) The Structure of Marketing Competition, Chikura Shobo [in Japanese].

Markus, M. L., & Benjamin, R. I. (1997). The magic bullet theory in IT-enabled transformation. MIT Sloan Management Review.

Palos-Sanchez, P., Saura, J. R., & Martin-Velicia, F. (2019). A study of the effects of programmatic advertising on users' concerns about privacy overtime. Journal of Business Research, 96, 61-72.

Pariser, E. (2011) The Filter Bubble: What the Internet Is Hiding from You. Penguin Press.

O'neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Weinmann, M., Schneider, C., & Brocke, J. V. (2016). Digital nudging. Business & Information Systems Engineering, 58, 433-436.

IAS [Integral Ad Science] (2023) Media Quality Report.

HIGHLIGHTING ETHICAL DILEMMAS IN SOFTWARE DEVELOPMENT: A TOOL TO SUPPORT ETHICAL TRAINING AND DELIBERATION

Pak Hei Li, Dharini Balasubramaniam

University of St Andrews (United Kingdom)

issacli7401@gmail.com; dharini@st-andrews.ac.uk

EXTENDED ABSTRACT

Introduction

Ethics is an increasingly important facet of software development and a key concern for all stakeholders of software systems (Gogoll et al., 2021). The now pervasive use of software in our daily lives means that a lack of ethical deliberation by software professionals can have profound consequences for stakeholders including end users. This recognition has led to the definition of codes of ethics for software engineering (SWECOE) (Gotterbarn et al., 1997) and sets of ethical principles applicable to specific areas such as artificial intelligence (AI) (Floridi & Cowls, 2019; Lo Piano, 2020). These codes and principles are intended to serve as a foundation for ethical decision-making by software engineers. However, software professionals are not typically offered training in ethical deliberation, or given pragmatic tools to support it. Therefore, they may find it challenging to apply the principles in practice (Hagendorff, 2020). The proliferation of project management tools, coupled with the increasing adoption of agile development methodologies, necessitates the early and systematic incorporation of ethical consideration into the development process. Our work aims to create a project management tool that helps to highlight ethical dilemmas during software development and integrates an extensible training resource for ethical deliberation. This paper describes the context, design, implementation, and evaluation of a proof-of-concept tool developed for this purpose and outlines avenues for further work.

Related work

The ACM / IEEE Software Engineering Code of Ethics, published in 1997, provides ethical guidance for software engineering professionals in the form of eight principles relating to: Public, Client and Employer, Product, Judgement, Management, Profession, Colleagues and Self (Gotterbarn et al., 1997). Several ethical principles and frameworks for AI have also been defined (Floridi & Cowls, 2019; Lo Piano, 2020; Prem, 2023). However, there has been limited research on how ethical deliberation can be supported during the creation of software artefacts (Gogoll et al., 2021) and whether effective ethics frameworks exist for software engineering processes in the industry (Mitchell et al., 2022).

Several solutions for facilitating ethical practices in software engineering have been proposed. A robust ethical framework helps in creating educational resources and training modules for students so that they develop into ethically aware professionals. Taherdoost et al. (2011) suggest including topics like “computer crime, privacy, intellectual property, accuracy, accessibility, morality, and awareness” in computer ethics courses. Additionally, the Ethical-Driven Software Development Framework (Lurie & Mark, 2016) encourages consideration of

ethics throughout the Software Development Life Cycle (SDLC), particularly in agile software development. Despite the significant work already done on defining codes and principles of ethics and ethical frameworks, applying these principles and frameworks in the workplace can be challenging (Mitchell et al., 2022) since there is still a lack of practical support for ethical deliberation in software engineering.

Methodology

A survey of related work was conducted to ascertain the state of the art in the area. Based on the findings, a prototype agile project management tool, incorporating support for ethics training and highlighting ethical dilemmas, was developed. The development itself followed an agile methodology. The prototype tool was evaluated to assess its usability and effectiveness in raising awareness of ethical dilemmas and refined to reflect the results of the evaluation. Ethics approval for the evaluation process was obtained from the authors' higher education institution.

Design, implementation and evaluation

The design and implementation phase focused on creating a proof-of-concept web-based ethics-centred project management tool aimed at software professionals. The features supported by the tool include an interactive ethics training resource, a Kanban project management board as an exemplar of agile project management, ethical framework infographics, ethics regulation checklists, a text adventure game, chatbot recommendations, ethics keyword flagging, and ethics self-assessment.

The resource illustrates the principles of the ACM Code of Ethics interactively via text adventure games and includes a recommendations component to extend users' knowledge of ethics (Figure 1). The project management tool highlights ethical concerns and resolutions to ethical dilemmas in decision-making. The highlighted dilemmas are related to the features or tasks in the product backlog (Figure 2). The Kanban board support the management of agile software development. Tasks are visualised on the board, allowing developers and project managers to see the state of each task at any time to promote ethical awareness and foster a culture of ethical responsibility and accountability throughout the software development process (Figure 3).

A client-server architecture was used for the web application system. The client handles the user interface for all the features listed above. The server manages backend processes, data processing, and business logic, and provides APIs for the client. The MERN stack (MongoDB, Express.js, React.js, Node.js) was used for the development of the tool.

A software artefact evaluation questionnaire was used to assess the software's effectiveness in supporting ethical practices in software development. Feedback was gathered from 21 participants in a higher education setting through opportunistic sampling. All participants were experienced in software development and used the software prototype before completing the user evaluation questionnaire. The questionnaire contained 19 statements which evaluated the interface clarity, ease of use, progress tracking, and integration of ethical guidelines of the tool during software development. It also assessed risk identification, resource support, and integration with existing project management tools. Participants were instructed to indicate whether they agreed or disagreed with each statement and to what extent.

Figure 1. In the ethical dilemma text adventure game, users engage with immersive scenarios, prompted by a graphical illustration. They select ethical practices from action buttons, indicated by changing button colours.

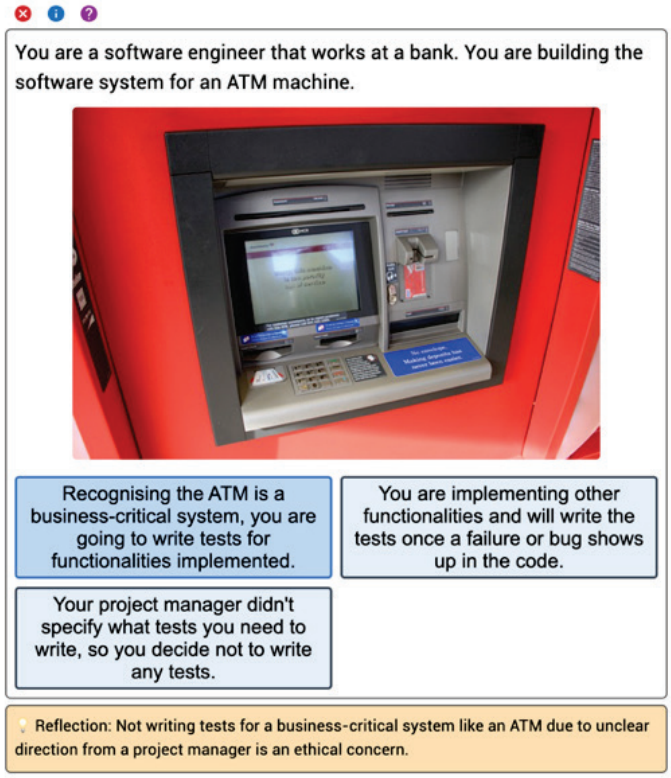


Figure 2. The task modal on the project board displays ethical principle tags, creation date, description editor, ethics flagging system, and chatbot recommendations.

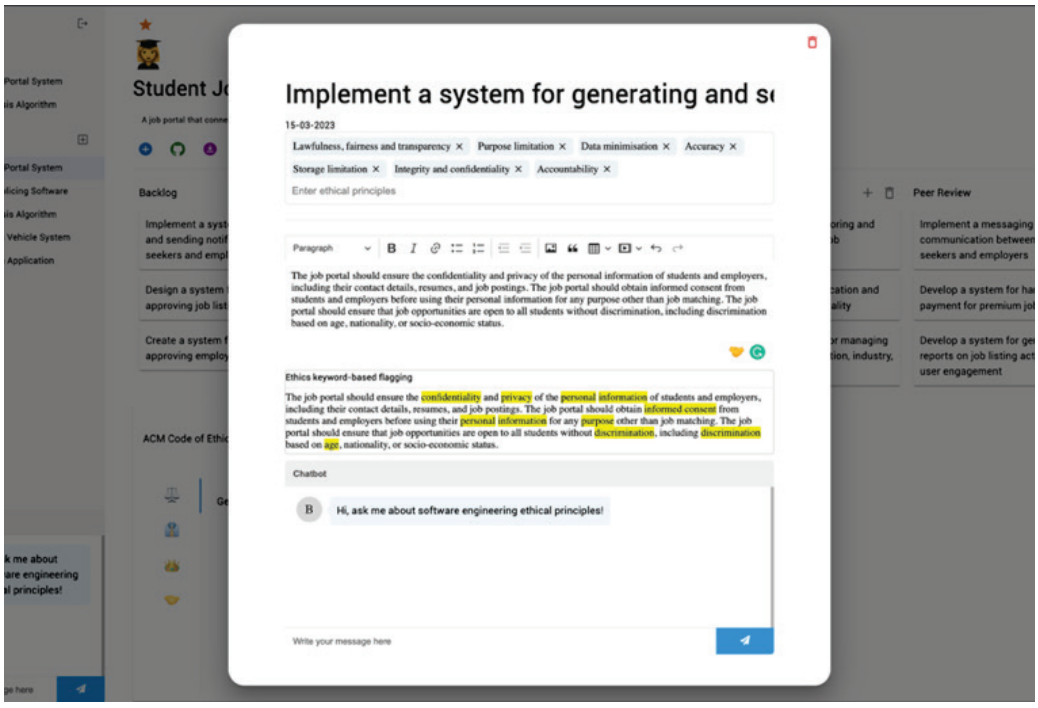
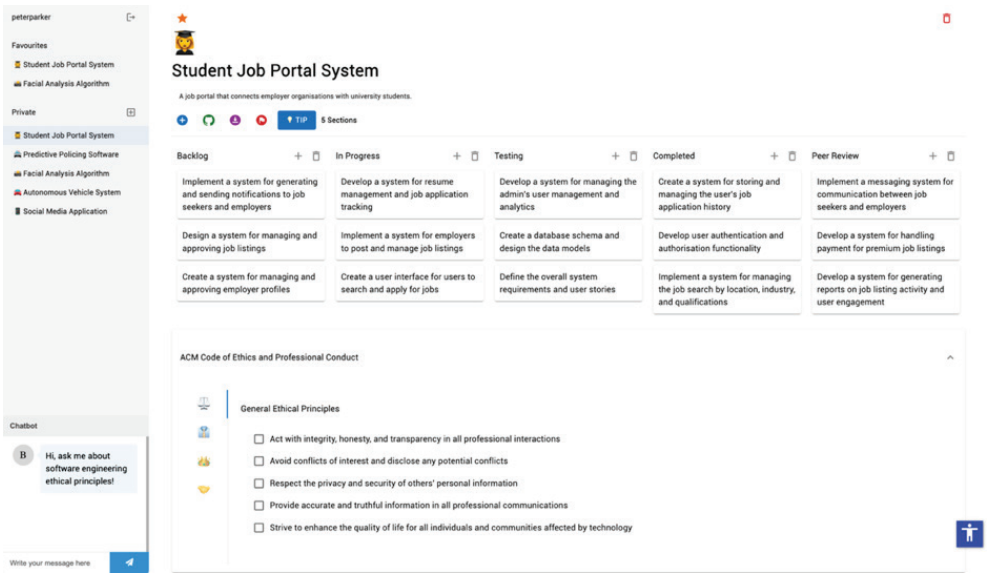


Figure 3. A sample project board for a student job portal system, with populated sections and tasks. The grey icons allow adding tasks or deleting sections. The red bin icon deletes the entire project board.



Results and discussion

The results from the evaluation provided valuable insights into the tool's usability and its effectiveness in promoting ethical awareness among users. For example, 66.67% of users "strongly agree" and 33.33% of users "somewhat agree" with the statement: "the tool supports the training and education of software developers on ethical dilemmas and best practices of software development ethics". Participants provided suggestions for improving functionality and user-friendliness.

According to participants, one key strength of the tool is its customisability, supporting the training of software developers on ethical issues and best practices. The ethical dilemma scenarios and adventure games were deemed insightful and engaging, making the learning process interactive. Additionally, the ethics checklist effectively assesses ethics practices in projects, ensuring ethical considerations in software design.

In summary, the user feedback highlights advantages such as customisability, effectiveness in assessing ethics practices, and interactive features. The suggested improvements include expanding the ethics glossary, enhancing the recommendation system, and implementing version control and peer review for ethical practices. Users would also like to have support for other development methodologies in addition to Kanban. Addressing these areas will enhance the tool's applicability, effectiveness, and user-friendliness.

Conclusion

This work contributes to the software engineering community by providing an extensible training resource and a custom agile project management tool highlighting ethical dilemmas. The resources fill the gap in ethical training and tools for software professionals. By raising awareness and providing necessary resources, this project improves ethical practices in the software development community.

KEYWORDS: Ethical dilemmas, software development, ethical training.

REFERENCES

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3831321

Gogoll, J., Zuber, N., Kacianka, S., Greger, T., Pretschner, A., & Nida-Rümelin, J. (2021). Ethics in the Software Development Process: From Codes of Conduct to ethical deliberation. *Philosophy & Technology*, *34*(4), 1085–1108. https://doi.org/10.1007/s13347-021-00451-w

Gotterbarn, D., & Miller, K. W. (2009). The public is the priority: Making decisions using the Software Engineering Code of ethics. *Computer*, *42*(6), 66–73. https://doi.org/10.1109/mc.2009.204

Gotterbarn, D., Miller, K. W., & Rogerson, S. (1997). Software engineering code of ethics. *Communications of the ACM*, *40*(11), 110–118. https://doi.org/10.1145/265684.265699

Gotterbarn, D. (2002). Software engineering ethics. *Encyclopedia of Software Engineering*. https://doi.org/10.1002/0471028959.sof314

Hagendorff, T. (2020). The ethics of AI Ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Lo Piano, S. (2020). Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, *7*(1). https://doi.org/10.1057/s41599-020-0501-9

Lurie, Y., & Mark, S. (2016). Professional Ethics of Software Engineers: An ethical framework. *Science and Engineering Ethics*, *22*(2), 417–434. https://doi.org/10.1007/s11948-015-9665-x

Mitchell, A., Balasubramaniam, D., & Fletcher, J. (2022). Incorporating ethics in software engineering: Challenges and opportunities. *2022 29th Asia-Pacific Software Engineering Conference (APSEC)*, 90–98. https://doi.org/10.1109/apsec57359.2022.00021

Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00258-9

Taherdoost, H., Sahibuddin, S., Namayandeh, M., & Jalaliyoon, N. (2011). Propose an educational plan for Computer Ethics and Information Security. *Procedia - Social and Behavioral Sciences*, *28*, 815–819. https://doi.org/10.1016/j.sbspro.2011.11.149

RESEARCH ETHICS FRAMEWORKS FOR ARTIFICIAL INTELLIGENCE: THE TWOFOLD NEED FOR COMPLIANCE REQUIREMENTS AND FOR AN OPEN PROCESS OF REFLECTION AND ATTENTION

Anais Resseguier

Trilateral Research (Ireland)

Anais.resseguier@trilateralresearch.com

EXTENDED ABSTRACT

This paper proposes to enhance research ethics frameworks for research projects developing and/or using Artificial Intelligence (AI). It highlights that these frameworks need both (a) requirements for compliance with emerging ethical and legal norms to govern this technology and (b) an open process of reflection and attention to research and innovation in this area. The field of AI ethics has seen intense developments since 2015 with numerous governmental and international bodies, institutions and companies creating guidelines, frameworks, and sets of principles for AI governance. Jobin et al. (2019) have analysed 84 of these documents and point to significant convergence on key principles, in particular transparency, justice and fairness, non-maleficence, and responsibility. However, these initiatives have also received sharp critiques from experts in the field, including that of being a form of "ethics washing" (Wagner, 2018; Resseguier and Rodrigues, 2020) or of reproducing existing power structures and inequalities (D'Ignazio and Klein, 2020). The proposed approach seeks to address these critiques by focusing on review processes as handled by research ethics committees (RECs), also called institutional review boards (IRBs). As the AI ethics field is currently working toward its operationalisation, research ethics constitute a powerful, but so far underdeveloped framework to make AI ethics more effective at the level of research (Santy et al. 2021).

A two-pronged approach to the operationalisation of AI ethics

The present paper proposes a two-pronged approach to the operationalisation of AI ethics in research ethics frameworks: (a) compliance with requirements imposed on researchers and (b) an open process of attention and reflection. In the words of the philosopher George Canguilhem, while the former aspect of ethics is about engaging with the norms, the second one attends to the *capacity* to determine the norms, i.e., the "normative capacity" (Canguilhem, 1991). Before presenting what this means concretely for AI research ethics (section 2), this paper makes a detour by the theory of ethics (section 1). It does so by drawing from works by Gertrude E.M Anscombe (1958) and Charles Mills (2005) that help provide conceptual clarity on the notion of ethics used primarily in AI ethics since around 2015 and ways to avoid critical pitfalls of this approach. It shows indeed how a clarification between the level of the norms (a) and that of the open process of reflection and attention (b), i.e., the "normative capacity" in Canguilhem's terms, helps to lift a confusion in AI ethics, a confusion that has weakened its potential effectiveness and has led to a number of legitimate critiques.

Compliance requirements and the potential role of the European AI Act

In the second section, this paper formulates a series of concrete recommendations for AI research ethics, based on the conceptual framework identified in the first section. To begin with, this paper encourages the imposition of particular requirements within research ethics frameworks embedded in institutions. This corresponds to the side of the norms requiring compliance (a) as identified in the model described in the first section. These norms, principles, or requirements, should be accompanied by mechanisms to ensure compliance, such as through the possibility of withdrawing funding if these are not fulfilled (this is for instance the case with the ethics appraisal scheme for research projects funded under the Horizon Europe Funding Program of the European Commission). Requiring compliance with certain criteria allows to put red lines and better orient AI research in a way that avoids potential harms caused by this technology, such as mass surveillance or discrimination. In this sense, research ethics takes the shape of "soft law" requiring compliance with certain obligations. This would help address the critique AI ethics has received of being "toothless", a form of "ethics washing", due to the absence of enforcement mechanisms.

Requirements from the European Union's AI Act currently under development will assuredly constitute a key reference for research ethics norms. Although, in the current form of the draft (as of June 2023), the obligations of the AI Act do not apply to scientific research, it is most likely that these obligations will nonetheless have a strong impact on AI research considering the need to anticipate placement on the market or to test in real world conditions (European Parliament, 2023). This paper explores the implications of the AI Act for research ethics frameworks and especially what the legal obligations in this regulation will mean for research ethics requirements and mechanisms to ensure compliance with these. In particular, it will investigate what the risk-based approach in the AI Act implies for research ethics and how to ensure compliance with the obligations at the different risk levels.

An open process of reflection and attention

In addition, ethics review frameworks offer a space for an open process of reflection and attention (b). The focus here is on questioning established norms and ways of doing through an open reflection and a continuously renewed form of attention to both technical advances in the field and social developments and concerns. This corresponds to the level of the "normative capacity", to use Canguilhem's terms as defined in the first section, i.e., the capacity to pay attention to the new situation, reflect on it, and challenge existing norms if needed to best adapt to the novelty one faces. Considering the uncertainty AI brings to societies, this constantly renewed attention and reflection is essential. For instance, in-depth critical social science and humanity (SSH) studies are crucial to engage such open reflection and renewed attention (e.g., Crawford, 2021). The submission of a societal impacts statement as part of an ethics submission for AI research projects can serve to embed such reflection within the ethics review process (Bernstein et al., 2021; Ada Lovelace Institute, 2022). Another option would be to carry out discussions with an expert on the ethical and social impacts of the AI system under development at several stages of the research project development. Strengthening the open process of reflection and attention at the research ethics level would help address the critique made toward AI ethics according to which it would fail to address structures of power and inequalities.

By distinguishing the level of the norms and that of the open process of attention and reflection, highlighting their respective values, and the way they relate to each other, this paper contributes

to advancing further ai ethics through its operationalisation in research ethics frameworks. The aim is eventually to make ai ethics more effective but also more thoughtful.

KEYWORDS: AI ethics; Research ethics review; AI Act; Ethics washing; Research ethics committees.

REFERENCES

Ada Lovelace Institute. (2022, Dec). Looking before We Leap. Expanding Ethical Review Processes for AI and Data Science Research. Retrieved from https://www.adalovelaceinstitute.org/report/looking-before-we-leap/

Bernstein, M. S., Levi, M., Magnus, D., Rajala, B. A., Satz, D., Waeiss, Q. (2021 Dec). Ethics and Society Review: Ethics Reflection as a Precondition to Research Funding. *Proceedings of the National Academy of Sciences* 118(52).

Canguilhem, G. (1991). *The Normal and the Pathological*. Translated by Carolyn R. Fawcett. Princeton: Princeton University Press.

Crawford, K. (2021) *Atlas of AI*. New Haven & London: Yale University Press.

D'Ignazio, C., Klein L. F. (2020). *Data Feminism,* Cambridge, MA; London, England: MIT Press.

European Parliament (2023, June), Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)) https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

Jobin, A., Ienca M., Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1(9), 389–99.

Mills, C. (2005). "Ideal Theory" as an Ideology. *Hypatia*, 20(3), 165–84.

Santy, S., Rani, A., & Choudhury, M. (2021). Use of Formal Ethical Reviews in NLP Literature: Historical Trends and Current Practices. *CoRR*, *abs/2106.01105*.

Rességuier, A., Rodrigues, R. (2020). AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics. *Big Data & Society* 7(2).

Wagner, B. (2018). Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping. In *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, ed. Emre Bayamlioglu et al., Amsterdam: Amsterdam University Press.

CLOSING THE AI RESPONSIBILITY GAP WITH THE CODE OF ETHICS

Don Gotterbarn, Marty J. Wolf

East Tennessee State University (United States), Bemidji State University (United States)

gotterba@gotterbarn.com; mjwolf@bemidjistate.edu

EXTENDED ABSTRACT

Like many types of technology, early computing technology was developed, at least in part, to support and advance military goals. For some, this put the technology on morally shaky grounds. As computing technology advanced and became more generally accessible, it increasingly was used by some to intentionally cause harm outside the military realm. Others used the technology for seemingly innocuous purposes and caused harm that was unseen by them, but felt by others. These situations sometimes led to an "ethical hysteria" where people used ethical expressions in unhelpful ways or developed tools that only offered partial help or, worse, went down the wrong path.

An early attempt to address harm caused by computing was the introduction of the notion of "Software Engineering" in 1968 at the first NATO Software Engineering Conferences. The goal was to indicate the professional technical skills that were needed to solve the "software crisis" of failed software (Naur & Randell 1968). The solutions tended to be technical and centered on the processes used to develop software; that software development process should follow an engineering model. This led to a report defining "software engineering techniques" to resolve the "software crisis." Over time this foundation was added to by the development of various software life cycles and developing software case tools were supposed to help create better, less harmful software. Yet, it wasn't until the 1990s that the IEEE-CS and the ACM publicly addressed the need for software engineering ethics standards in the IEEE/ACM Software Engineering Code of Ethics (Gotterbarn et al. 1997).

This is the start of a pattern (described in the full paper) whereby each new computing ethical awakening repeats mistakes from earlier eras. There is a pattern of concerns present in many cases reaching back to the first software crisis and now to AI. The most pressing concern stemming from these "crises" for computing ethics is: what structural support can best help those who want to use computing in positive ways and yet have difficulty doing so?

In this paper we focus on AI. There are a number of things that have changed since the last major ethical awakening in computing. First, and importantly, there is a broad range of people who are at the table discussing the ethics and social implications of AI. The different perspectives brought by philosophers, ethicists, linguists, sociologists, computer scientists, mathematicians, data scientists, humanities scholars and so many more, all come to bear on identifying actual and potential harms of computing technology. Further, they offer suggestions on different ways to prioritize those harms.

The other major change is that there is a greater misalignment between corporate interests and "good AI" (AI that does good for society) than with previous ethical awakenings. Developing software that met specification was good for business. Having developers understand the same software lifecycles added efficiencies to software development and contributed to the

corporate bottom line. These days, while products like ChatGPT may be good for OpenAI's bottom line, there is clear harm being caused to other businesses, to education, and even to democracy itself, and this is on top of harms caused by its development (see Bender et al. 2021).

With experts from these different fields coming to bear on AI, significant new problems have been identified, including biased decisions, misclassification, overgeneralizations, lack of contextual understanding, adversarial attacks, and unintended consequences. When a person is responsible for these kinds of problems, they are held accountable or blamed as the cause. When these judgments are left to an AI system there is a difficulty assigning responsibility for problems or bad decisions. Adreas Matthias has called this situation where no human can be morally responsible or liable for a machine's behavior the "responsibility gap" (2004).

Many such as (Goetze 2022, Kiener 2022, Rubel 2019, Santoni de Sio and Mecacci 2021, Tigard 2021) have addressed facets of the responsibility gap, including when and whether it exists. We consider one of them here and the others in the full paper. Munch et al. (2022) argue that there are times when the responsibility gap is good even in the absence of psychological dilemmas. They argue that holding a person responsible for wrong-doing causes that person some amount of harm. In situations where an automated system is equally as effective as a person in making decisions of consequence, no person comes to bear the harm associated with wrong-decision making when there is a responsibility gap. Our contention is that this position and other arguments surrounding the notion of responsibility gaps misdirect discussions about responsibility.

A major problem is that some of the discussion about the responsibility gap has focused on a limited sense of responsibility, one related to blame in some form. This same sense of responsibility was used during the software crisis of the 1960s to blame developers for the failure to develop reliable systems. The result was a system that emphasized finding a program's errors rather than people learning how to take action to decrease the risk of similar errors in future programs. Further, blame would frequently be passed to the client for inadequately specifying requirements. Responsibility for the moral issues surrounding the requirements and the way a system developed were not considered. There was significant effort to develop precise technical requirements as a problem solution.

John Ladd called this responsibility "negative responsibility," a responsibility assigned after the fact. It primarily tries to excuse people from moral responsibility, a legal search for extenuating circumstances, for example. He champions a positive sense of responsibility for what ought to be done. Unlike negative responsibility which tends to be direct, positive responsibility can be indirect. Ladd argues that pointing to the technology does not remove this sense of positive responsibility. Positive responsibility engages with the prospect that things might happen. Guidance from Principle

2.2 of the ACM Code of Ethics and Professional conduct is clear: "Professional competence starts with technical knowledge and with awareness of the social context in which their work may be deployed." Computer professionals are responsible for applying standards within their profession and attempting to avoid anticipatable negative ethical impacts of their work.

The AI responsibility gap discussion misses this opportunity to engage with positive responsibility. Underlying the AI responsibility gap is an assumption of a causal chain looking for a particular event. Implicitly, this makes a standard responsibility denial move, appealing to the complexity of the system, much easier and misses an opportunity to change the behavior of the system's developer. For AI systems, an appeal to the responsibility gap can be used to justify the

development of systems that cause harm. (Google initially did this when Safiya Noble pointed out how their search completion algorithm reinforced racist stereotypes.) The advocacy/acceptance of such positions are inconsistent with ethical computing.

Professional responsibility also includes premeditated concern for the consequences of one's actions on others. This kind of approach is anticipated by the ACM Code of Ethics and Professional Conduct and advocated for by Goterbarn et al. (2022). We highlight another approach next that is applicable to AI and has AI developers focus on how the AI systems are built and how decisions are made by AI developers.

One of the authors participated in the development of several international police criminal intelligence systems. Ethical issues were considered and mitigated in the design of the projects rather than after the systems were built. Implicit and explicit values in design choices and the intentional and unintentional value choices made in technology development were made explicit.

The project set up design guidelines so that solutions to issues were designed before the system was implemented. This helped to change the focus from an overview of ethics (e.g. informed consent) to a deeper focus on the technologies, their impact on society, and the ethical issues that the different technologies may raise.

For example, in order to identify bias in decisions and inferences made from the data, the design process included transparency tools such as understandable process logs and logging mechanisms that tied change details to a user. To ensure the integrity of the data and increase the reliability of inferences made from it, the design included a 'reliability tag' attached to all data.

Not all AI systems make bad decisions, and many are designed to mitigate risks through rigorous testing, validation, and ongoing monitoring. While these after the fact tools are important, positive responsibility calls for more. Using tools like the ACM Code of Ethics and Proactive CARE (Gotterbarn et al. 2022) is essential to ensure responsible and ethical AI deployment. The Code of Ethics provides guiding principles that lead to better design decisions and help developers use positive responsibility to reduce or even eliminate the AI responsibility gap.

KEYWORDS: Responsibility, Responsibility Gap, Artificial Intelligence Responsibility, ACM Code of Ethics.

REFERENCES

ACM Code of Ethics and Professional Conduct (2018). https://www.acm.org/code-of-ethics

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

Goetze, T. (2022). Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement. FAccT '22.

Gotterbarn, D. (2001). Informatics and professional responsibility. Science and Engineering Ethics, 7, 221–230.

Gotterbarn, D., M.S. Kirkpatrick, and M.J. Wolf. (July, 2022). ”From the page to practice: Support for computing professionals using a code of ethics,” ETHICOMP 2022.

Don Gotterbarn, Keith Miller, and Simon Rogerson. (1997). Software engineering code of ethics.

Commun. ACM 40, 11 (November 1997), 110-118. http://doi.org/10.1145/265684.265699

Kiener, M. (2022). Can we Bridge AI’s responsibility gap at Will?. Ethic Theory Moral Prac 25, 575–593. https://doi.org/10.1007/s10677-022-10313-9

Ladd, J. (1988). Computers and Moral Responsibility: A Framework for an Ethical Analysis, in: Gould, Carol (ed.) The Information Web: Ethical and Social Implications of Computer Networking, Westview Press.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6, 175–183. https://doi.org/10.1007/s10676-004-3422-1

Munch, L., Mainz, J. & Bjerring, J.C. The value of responsibility gaps in algorithmic decision-making. Ethics Inf Technol 25, 21 (2023). https://doi.org/10.1007/s10676-023-09699-6

Naur, P. & Randell, B., eds. (1968). Software Engineering: Report on a conference sponsored by the NATO Science Committee. http://homepages.cs.ncl.ac.uk/brian.randell/NATO/nato1968.PDF

Rubel, A., Castro, C. & Pham, A. (2019). Agency Laundering and Information Technologies. Ethic Theory Moral Prac 22, 1017–1041. https://doi.org/10.1007/s10677-019-10030-w

Santoni de Sio, F., Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. Philos. Technol. 34, 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Tigard, D.W. (2021). There Is No Techno-Responsibility Gap. Philos. Technol. 34, 589–607. https://doi.org/10.1007/s13347-020-00414-7

ON THE ETHICS OF MISAPPLYING A CODE OF ETHICS

Marty J. Wolf, Catherine Flick

Bemidji State University (United States), De Montfort University (United Kingdom)

mjwolf@bemidjistate.edu; cflick@dmu.ac.uk

EXTENDED ABSTRACT

The computing profession sits apart from other disciplines in academia, in science and engineering, and in industry. It is not a profession in the strict sense, in the sense that engineers, doctors, or lawyers are. There is no licensure. There is no board to which ordinary citizens can make claims of malpractice that can strip credentials from someone who performs their job poorly or unethically. Regardless, there is a deep and mutually beneficial relationship among academia, computing companies, and conferences that is rare in other disciplines. The primary professional organizations play a key role in this relationship, bringing industry and academia together. These organizations sponsor conferences and are the leading publishers of computing-related research.

Each of two large professional organizations that represent computing professionals has its own code of ethics. The ACM Code of Ethics and Professional Conduct is longer and provides more guidance on its principles (ACM 2018). The IEEE Computer Society (IEEE-CS), the part of IEEE focused on computing professionals, does not have its own code of ethics, but its members are subject to the IEEE Code of Ethics (IEEE 2020), which is shorter than the ACM's and fits comfortably on a single sheet of paper. For the purposes of this paper these codes share two important features: the codes are directed only at individuals and those individuals are members of the respective professional organization.

IEEE-CS claims to have approximately 375,000 "community members" who represent 168 countries worldwide. ACM claims to have approximately 100,000 members, about half from North America and half from the rest of the world. While it is reasonable to assume that there are computing professionals who belong to both organizations, it is safe to say that combined, they represent no more than half a million computing professionals worldwide. A further point that lends to the importance of the considerations in this paper is that IEEE-CS does not appear to be a strong promoter of its code of ethics. There is no mention of its code of ethics on its landing page or its "about" page. Thus, there is a question about how actively it promotes its code of ethics. This observation is not meant to be a criticism of IEEE-CS, but rather, it is intended to help motivate the point that relatively few computing professionals may even have knowledge of the fact that they are subject to IEEE's code of ethics.

The ACM is upfront about the expectation that members agree to abide by their Code of Ethics and Professional Conduct on the membership application page. Further, ACM is clear to anyone applying for membership about its dedication to "promoting the highest professional and ethical standards." ACM expects its members to share that value, and ties the requirement to abide by the Code of Ethics and Professional Conduct to membership in the organization. Additionally it has a thorough and publicly available complaints handling process for suspected violations which can potentially result in membership being revoked, being banned from publishing in ACM publications and attending ACM events including conferences.

This leads us to the following concern: Given that so many computing professionals are not part of an international professional organization that holds as a key value the highest professional and ethical standards, should a professional organization such as ACM or IEEE-CS hold computing professionals generally to such a standard, and if so, how should it go about ensuring that all computing professionals are held to that standard?

Since IEEE and ACM are particularly powerful forces in computer science publishing, one option would be to expand the range of applicability of their respective codes to those who publish in their journals. This would serve two purposes. First, it would give an avenue to better educate computing professionals about the professional responsibilities found in the codes of ethics. Second, it would increase the range of sanctions that might be applied in cases where a violation is found. The possibility of losing publishing privileges in some or all of IEEE's or ACM's journals can be impactful, especially for academics. Unfortunately, this line of thinking only addresses a subset of computing professionals.

We recognize that there are many country-based professional organizations that do promote high ethical standards. There is a question about whether codes of ethics such as IEEE's and ACM's are truly reflective of international values as their development was done exclusively in English. Work done by Shannon Vallor suggests, however, that the values reflected in these codes of ethics may indeed be shared more globally (2016). The full paper addresses the appropriateness of international organizations collaborating with national or local professional organizations in developing, promoting, and applying codes of ethics to those entities where a given code of ethics may not be designed to apply.

A second concern we will address in the paper is that these professional codes of ethics do not apply to groups--and in particular companies. This concern manifests itself in a number of ways.

First, critiques of a code may misapply the code. For example, in a commentary where Aaditeshwar Seth called for the ACM Code of Ethics to "embrace goals such as achieving equality and overturning unjust social and economic structures through technological inventions," they identified a shortcoming of the code by providing an examples of corporate and government failures to recognize goals that are harmful to people and society. The code in this case was used for a purpose that it was not designed for.

A second concern that is sometimes raised with these professional organizations is that they do not apply their code of ethics to the companies that are developing technology. This criticism is particularly poignant when an organization such as ACM expresses its dedication to "promoting the highest professional and ethical standards" in public ways and expects its members to uphold those same standards. Even should a code of ethics be applied to a tech company, the process for investigating a complaint is unclear. Imagine that ACM tried to investigate a major company such as Alphabet for YouTube's recommendation system, which tends to lead people to some of the most extreme content on the site. How would such an investigation be carried out? Who would do the interviews? Would people at the tech company be allowed to talk without fear of retribution from their employer? What are reasonable sanctions should the company be found to have violated the code of ethics?

Organizations such as ACM and IEEE are certainly large enough and powerful enough to make public statements (as a possible sanction) about the harms caused by a tech company's product or actions. There is every reason to expect some sort of retaliation, though, due to how generously major tech companies support ACM and IEEE conferences through their financial and in-kind

contributions. Such a sanctioning regime may create a two-tiered system where companies that are supportive of conferences are less likely to face scrutiny than those that do not.

The IEEE-CS program that allows corporate memberships may be the seed of an approach to address these concerns. A corporate membership that came only with a commitment to the highest professional and ethical standards may provide a foundation for a "name and praise" system, rather than a "name and shame" system. A name and praise system would identify companies that adhere to best practices for ethical computing as well as techniques for verifying that those practices are actually effective. Yet there are limitations to this approach as well.

This paper will develop these questions and suggest some ways forward in order to foster a lively discussion about application - and misapplication - of codes of ethics.

KEYWORDS: Code of Ethics, Professionalism, Computing Profession, Professional Organizations, Professional Organization Responsibilities.

REFERENCES

ACM Code of Ethics and Professional Conduct (2018). https://www.acm.org/code-of-ethics

IEEE Code of Ethics (2020). https://www.ieee.org/content/dam/ieee-org/ieee/web/org/about/corporate/ieee-code-of-ethics.pdf

Aaditeshwar Seth. (2023). What's Missing in the ACM Code of Ethics and Professional Conduct. interactions 30, 3 (May + June 2023), 44–47. https://doi.org/10.1145/3588003

Vallor, S. (2016). Technology and the Virtues. New York: Oxford University Press.

SUSTAINABLE SUCCESS: UNRAVELING THE RELATIONSHIP BETWEEN CSR INITIATIVES, HAPPINESS, AND PURCHASE INTENTION IN FASHION RETAILERS ACROSS CHANNELS

Pablo Gutiérrez-Rodríguez, Pedro Cuesta-Valiño, Estela Nuñez-Barriopedro, Blanca García Henche

Universidad de León (Spain), Universidad de Alcalá (Spain), Universidad de Alcalá (Spain), Universidad de Alcalá (Spain)

pablo.gutierrez@unileon.es; pedro.cuesta@uah.es, estela.nunezb@uah.es, blanca.garcia@uah.es

EXTENDED ABSTRACT

In these times of uncertainty and economic crisis, one of the main concerns for business managers should be to stimulate the purchasing intention of their potential consumers in order to consolidate or increase their market share in the globalized market (Zhang and Ma, 2020). Faced with this business challenge, corporate management must design strategic marketing actions aimed at incentivizing their customers' need to purchase their products or services (Dash et al., 2021). As is well known, this behavior is influenced by multiple factors such as cultural, social, and psychological aspects (Zupan et al., 2023).

This research highlights the need to delve into the understanding of perceived dimensions of CSR and their potential relationships with happiness and purchase intentions of fashion consumers in both physical store and digital environments. Therefore, the main objective of this research is to determine how the dimensions of CSR - economic, legal, ethical, and philanthropic - influence happiness and purchase intentions in the fashion sector.

Corporate Social Responsibility (CSR) is a concept that integrates marketing activities with a social focus, as supported by research (Galbreath, 2010). These activities encompass environmental conservation, community investment, resource preservation, and altruistic contributions (Nejati et al., 2017). Graafland et al. (2004) propose a model that includes three dimensions: economic, social, and ecological. Similarly, Carroll's model (1979, 1991; Carroll & Brown, 2018) delineates four dimensions: economic, legal, ethical, and philanthropic. Social responsibility can be defined as the extent to which companies assume different types of responsibility towards their stakeholders (Carroll, 1979, 1991; Carroll and Brown, 2018; Maignan, 2001; Park et al., 2014).

Consumer happiness has gained significant relevance in studies conducted by positive psychology and is increasingly being studied within business and marketing disciplines (Kennison, 2022; Pipoli de Azambuja et al., 2022). Happiness is often conceptualized as subjective well-being, quality of life, or pleasure (Ravina-Ripoll et al., 2022; Singh et al., 2022). Happiness can be categorized into two major dimensions: hedonic happiness and eudaimonic happiness. Hedonic happiness refers to the experience of pleasure or subjective well-being in the absence of pain and stress, while eudaimonic happiness encompasses the pursuit of a meaningful and fulfilling life through personal growth, excellence, and the realization of one's potential (LeFebvre and Huta, 2021).

Purchase intention is a key focus in this study, aiming to identify factors that influence consumers' intent to purchase through an in-depth analysis of relevant literature. Previous research has examined the impact of consumer characteristics, merchant and product characteristics, and the type of products on purchase intention (Chang et al., 2022; Morwitz, 2014; Shwu-Lng and Chen-Lien, 2009; Shah et al., 2012). Purchase intentions are often used to predict sales (Morwitz et al., 2007).

The research you described is a cross-sectional descriptive study that relies on primary data collected through a questionnaire. The target population for the study is the Spanish population aged 16 to 64. The data collection period spanned from May to August 2022. A total of 1,296 valid questionnaires were collected from the participants.

Survey design and sample size. The sample size of 1,296 respondents provides a margin of error of +/-2.78% with a 95.5% confidence interval, assuming that the proportion of the population with a particular characteristic (p) is 0.5 and the proportion without that characteristic (q) is also 0.5. The margin of error indicates the range within which the true population parameter is expected to fall.

It's worth noting that the confidence level, margin of error, and assumptions about p and q can affect the precision and reliability of the study's findings. The chosen sample size appears to provide a reasonably representative sample of the Spanish population for the purpose of the research. Table 1 likely provides additional information related to the sample and its characteristics.

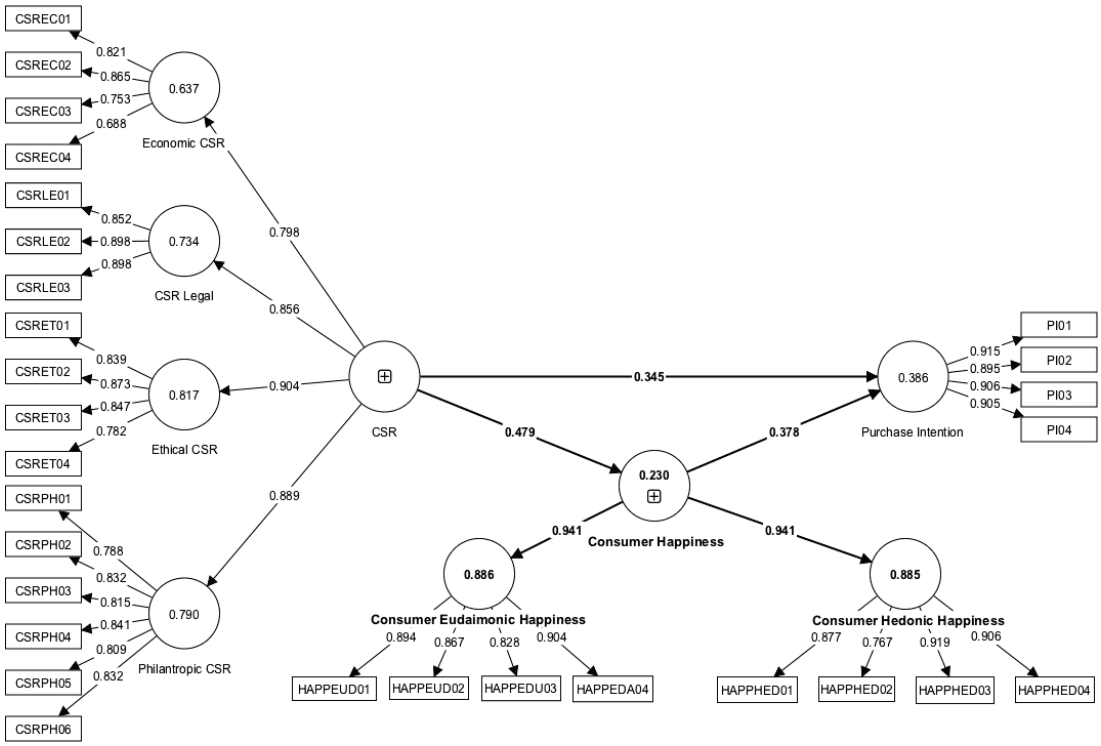
Table 1. Technical datasheet

Universe	**Males and Females Aged 16–64**
Geographical scope	Spain
Fieldwork	From May to August 2022
Sampling	Discretionary non-probabilistic by quotas
Sample	1,296 valid surveys
Sample error	+/−2.78 with a 95.5% confidence level and $p = q = 0.5$

The survey consists of two sections. The initial part focuses on capturing demographic characteristics and respondent behavior. The second section evaluates the proposed model's five dimensions using a 5-point Likert-type scale ranging from 1 ("strongly disagree") to 5 ("strongly agree"). To mitigate common method variance (CMV), this second part is organized by variable. The refinement process resulted in a final set of 29 items for the model.

Regarding CSR, the distribution included four items for economic dimension, three for legal dimension, four for ethical dimension and six for philanthropic dimension (Podnar and Golob, 2007; Perez and Bosque, 2014; Gunesh and Geraldine, 2015). For eudaimonic happiness, four items were utilized, and for hedonic happiness, five items were used (Fu and Wang, 2021; Cuesta et al., 2022). Lastly, four items were allocated to measure Purchase Intention (Duffett, 2015).

Figure 1. Results.



Findings. The measurement of CSR and Happiness involved assessing them as reflective second-order constructs. The assessment of reliability and validity indicates the robustness of the model's components. Additionally, it is crucial to examine the loadings of the dimensions within the two second-order variables. As presented in Table 2 and Figure 2, the indicators for CSR - Economic (0.82), Legal (0.86), Ethical (0.90), and Philanthropic (0.88) - as well as Happiness - Eudaimonic (0.93) and Hedonic (0.96) - demonstrate their effective representation of these variables across all dimensions (Figure 2).

The hypotheses collectively suggest a positive and significant relationship between CSR, Happiness, and purchase intention. It is posited that this relationship is meaningful and has a significant positive influence. The influence values of 0.47 on Happiness and 0.38 on purchase intention strongly support this proposition. Specifically, the coefficient of 0.35 indicates a substantial positive impact of CSR on purchase intention through Happiness.

In **conclusion**, this study contributes to the understanding of the impact of CSR and Happiness on purchase intention in an omnichannel environment, specifically within the fashion industry. While sustainability-based strategies are already prevalent in organizations and contribute to value creation, limited research has explored the influence of happiness and its implications for purchase intent in this context.

The findings confirm that CSR plays a pivotal role in improving Happiness and attracting consumers with high purchasing potential. However, the results also indicate that CSR requires a mediator to have a truly remarkable influence on consumers' purchasing intentions. Therefore, it is crucial for brands to adopt a sustainability-focused strategic approach that enhances customer experience and strengthens brand affinity across both digital and physical retail channels.

The proposed model emphasizes the importance of adopting a proactive and future-oriented strategic outlook, where companies strive to anticipate customer needs. These findings provide valuable insights for brand managers, retailers, and academics, serving as decision-making resources.

To enhance customer experience and foster customer loyalty, organizations should effectively manage CSR across all its dimensions, as indicated by the relevant results. The higher the level of happiness, stimulation, and autonomy experienced by the consumer, the stronger their affective experience in their relationship with the brand. Consequently, customer experience also yields other well-studied benefits, such as increased repeat purchases, willingness to recommend the service to others, and resistance to switching to competitors, all of which contribute to fostering customer loyalty.

KEYWORDS: CSR, Consumer Happiness, Omnichannel, Fashion, Retailers.

REFERENCES

Carroll, A. B. (1979). A three-dimensional conceptual model of corporate performance. *Academy of Management Review*, 4(4), 497–505.

Carroll, A. B. (1991). Corporate Social Performance Measurement: A Comment on Methods for Evaluating an Elusive Construct, in L. E. Post (ed.), *Research in Corporate Social Performance and Policy*, 12, 385–401.

Carroll, A. B., & Brown, J. A. (2018). Corporate Social Responsibility: A Review of Current Concepts, Research, and Issues. *Corporate Social Responsibility (Business and Society 360)*, Vol. 2, Emerald Publishing Limited, Bingley, (pp. 39-69).

Chang D., Lihang C., & Cuixia L. (2022). Can Information Intervention Enhance Consumers' Purchase Intentions of Organic Agricultural Products? A Choice Experiment Based on Organic Milk. *Journal of Healthcare Engineering*, 1256796.

Cuesta-Valiño, P., Gutiérrez-Rodríguez, P., & Núnez-Barriopedro, E. (2022). The role of consumer happiness in brand loyalty: a model of the satisfaction and brand image in fashion. *Corporate Governance*, 22(3), 458-473.

Duffett, R. G. (2015). Facebook advertising's influence on intention-to-purchase and purchase amongst Millennials. *Internet Research*, 25(4), 498-526.

Fu, Y. K., & Wang, Y. J. (2020). Experiential value influences authentic happiness and behavioural intention: lessons from Taiwan's tourism accommodation sector. *Tourism Review*, 76(1), 289-303.

Galbreath, J. (2010). Drivers of Corporate Social Responsibility: The role of formal strategic planning and firm culture. *British Journal of Management*, 21(2), 511–525.

Graafland, J. J., Eijffinger, S. C. W., & Smid, H. (2004). Benchmarking of Corporate Social Responsibility: Methodological problems and robustness. *Journal of Business Ethics*, 53, 137–152.

Gunesh, R. V., & Geraldine, R. W. (2015). Do CSR practices of banks in Mauritius lead to satisfaction and loyalty? Studies in Business and Economics, 10(2), 128-144.

Kennison, S. (2022). Humor and resilience: relationships with happiness in young adults. *Humor*, 35.

LeFebvre A. & Huta, V. 2021. Age and Gender Differences in Eudaimonic, Hedonic, and Extrinsic Motivations. *Journal of Happiness Studies*, 22(5), 2299-2321.

Maignan, I. (2001). Consumer Perceptions of Corporate Social Responsibility: A Cross Cultural Comparison. *Journal of Business Ethics*, 30(1), 57–73.

Morwitz, V. (2014). Consumers' Purchase Intentions and their Behavior, *Foundations and Trends® in Marketing*, 7(3), 181-230.

Nejati, M., Quazi, A., Amran, A., & Ahmad, N. H. (2017). Social responsibility and performance: does strategic orientation matter for small businesses? *Journal of Small Business Management*, 55, 43-59.

Park, J., Lee, H., & Kim, C. (2014). Corporate social responsibilities, consumer trust and corporate reputation: South Korean consumers' perspectives. *Journal of business research*, 67(3), 295-302.

Pipoli de Azambuja, G., Gustavo Rodríguez P., & Vargas, E. T. (2023) Marketing of Happiness: The Role of Customer Loyalty on Happiness. *Journal of Promotion Management*, 29(2), 228-258.

Ravina Ripoll, R., Galván Vela, E., Sorzano, M., & Ruíz-Corrales, M. (2022). Mapping intrapreneurship through the dimensions of happiness at work and internal communication. *Corporate Communications: An International Journal*, 28.

Shah, S., Aziz, J., Jaffari, A. R., Waris, S., Ejaz, W., Fatima, M., & Sherazi, S. (2012). The impact of brands on consumer purchase intentions. *Asian Journal of Business Management*, 4, 105-110.

Shwu-Ing, W. & Chen-Lien, L. (2009). The influence of core-brand attitude and consumer perception on purchase intention towards extended product. *Asia Pacific Journal of Marketing and Logistics*, 21, 174-194.

Singh, K., Bandyopadhyay, S., & Saxena G. (2022). An Exploratory Study on Subjective Perceptions of Happiness from India. *Frontiers in Psychology*, 13, 823496.

Pérez, A., & Rodríguez del Bosque, I. (2015). An integrative framework to understand how CSR affects customer loyalty through identification, emotions and satisfaction. *Journal of Business Ethics*, 129(3), 571-584.

Podnar, K. & Golob, U. (2007). CSR expectations: the focus of corporate marketing. *Corporate Communications: An International Journal*, 12(4), 326-340.

Zhang, Y., & Ma Z. F. (2020). Impact of the COVID-19 Pandemic on Mental Health and Quality of Life among Local Residents in Liaoning Province, China: A Cross-Sectional Study. International Journal of Environment Research and Public Health, 17(7), 2381.

Zupan, Z., Minyu, L., & Huijing, G. (2023). Exploring the Mechanism of Consumer Purchase Intention in A Traditional Culture Based on The Theory of Planned Behavior. *Frontiers in Psychology*,14, 315.

THE ETHICAL AND LEGAL CHALLENGES OF DATA ALTRUISM FOR THE SCIENTIFIC RESEARCH SECTOR

Ludovica Paseri

Law Department, University of Turin (Italy)

ludovica.paseri@unito.it

EXTENDED ABSTRACT

Scientific research nowadays is increasingly data-driven and therefore requires a growing amount of data, which need to be accessible and of high quality. The data altruism mechanism, that results as a means to meet this demand, is regulated in the Data Governance Act (DGA, hereinafter). The DGA is a Regulation of the European Union, which is applicable from 23 September 2023[8], that aims to "foster the availability of data for use by increasing trust in data intermediaries and by strengthening data-sharing mechanisms across the EU", as described in the explanatory memorandum accompanying the proposal for a Regulation[9]. The DGA is a crucial part of the so-called "politics of data" (Pagallo, 2022) developed by the European Commission in 2020[10] and can also be considered as complementary to the Open Data Directive (ODD, hereinafter)[11], integrating the European framework on data sharing and reuse (Ruohonen & Mickelsson 2023). The Article 3 of the DGA, which identifies the scope of application, expresses the complementarity between the DGA and the ODD by stating that the DGA provides for the reuse of certain categories of data, such as data held by the public sector that are protected on the basis of commercial confidentiality, statistical confidentiality, protection of third parties' intellectual property rights and protection of personal data. Therefore, the DGA concerns the reuse of those public sector data excluded from the scope of the ODD (Van Eechoud, 2021, p. 376).

Data altruism is defined by the DGA, in the Article 2(16):

> 'data altruism' means the consent by data subjects to process personal data pertaining to them, or permissions of other data holders to allow the use of their non-personal data without seeking a reward, for purposes of general interest, such as scientific research purposes or improving public services.

This contribution aims to investigate data altruism mechanism for the scientific research sector. This mechanism, based on the voluntary release of data, raises several ethical and legal

[8] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), ELI: http://data.europa.eu/eli/reg/2022/868/oj.

[9] Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act), COM/2020/767 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52020PC0767.

[10] European Commission Communication, A European strategy for data, COM/2020/66 final (2020), ELI: https://eur-lex.europa.eu/legal- content/EN/TXT/?uri=CELEX:52020DC0066.

[11] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast), ELI: http://data.europa.eu/eli/dir/2019/1024/oj.

challenges. From the legal viewpoint, the data altruism in the research sector raises the following challenges: (1) the risk of fragmentation; (2) the security concerns; and (3) the duty of control on data altruism organisations.

From the ethical perspective, there is (1) a problem of terminological uncertainty; (2) a need for decoding the concept of altruism underlying the release of data; and (3) a concern related to the autonomy of subjects in giving consent.

The mechanism of data altruism identifies the activities of several actors: (i) the data subject of personal data; (ii) the data holder of non-personal data; (iii) the data altruism organisations; (iv) the data users; and (v) the competent authority for registration. Underlying the operations of this multiplicity of actors are two conditions, expressed in the Article 2(16) of the DGA. The first condition is that reuse by data subjects and data holders must be granted to the data altruism organisation, free of charge, not in return for a reward. This condition may be interpreted as a measure to avoid the establishment of a buying and selling of personal data.

The second condition is that the reuse must pursue general interest purposes. These purposes are specified in the Recital 45, which states: "Such purposes would include healthcare, combating climate change, improving mobility, facilitating the establishment of official statistics or improving the provision of public services. Support to scientific research, including for example technological development and demonstration, fundamental research, applied research and privately funded research, should be considered as well purposes of general interest".

The mechanism of the data altruism, based on data "voluntarily made available by individuals or companies for the common good" (Proposal DGA, Explanatory Memorandum, 2020, p. 8), generates a considerable impact on the data management in the scientific research sector.

According to the European institutions, the data altruism mechanism involves an articulated process with several phases. First, any entity intending to be recognised as data altruism organisation has to undergo a registration process, and among other information, it is required to declare that "the purposes of general interest it intends to promote when collecting data" (Article 19(4)h, DGA). The general interest is primarily identified by scientific research by the DGA, when presenting the notion of data altruism, the Article 2(16) states that this mechanism shall be realised "for purposes of general interest, such as scientific research purposes or improving public services". However, it is often difficult to be precise about the aims pursued in a specific scientific research project (Pagallo & Bassi, 2013, p. 183). After that, if the requesting entity meets all the requirements laid down by the DGA, it will be included in the national register of data altruism organisations, by the competent national authority or authorities, within 12 weeks from the date of application, pursuant to the Article 19(5) of the DGA.

The voluntary release of personal data by data subjects, or non-personal data by data holders, to data altruism organisations is based on the provision of the consent. The consent needs to be given in compliance with the two conditions described above, i.e., no reward and public interest purposes. The registered data altruism organisations provide to several natural and legal persons the possibility to process the data they hold, for purposes of general interest, eventually on the basis of a fee. Each data altruism organisation is required to keep accurate records – very similar to the processing register set out in the Article 30 of the General Data Protection

Regulation (GDPR, hereinafter)[12] – concerning a set of accurate information about the specific data processing activities, based on the data altruism consent. In addition, each data altruism organisation, pursuant to the Articles 20 and 21 of the DGA, has several reporting obligations towards data holders. In particular, entities are required to communicate the purposes for which further processing of data is permitted to third parties. In this regard, it is significant that the corresponding article in the proposal of the Regulation (Article 19 of the DGA Proposal) made explicit reference to the duty to communicate "any processing outside the Union", expression that no longer appears in the wording of the Regulation. In addition, it is relevant (and problematic) that any organisation of data altruism also has a function of control over the entire lifecycle of the data that is given to third parties to process. The Article 21 of the DGA states, in fact, that the "entity shall also ensure that the data is not be used for other purposes than those of general interest for which it permits the processing". However, the problematic aspect emerging in this phase is represented by the fact that the purpose and regulation of reuse of public sector data is intrinsically generic and open to any possible use of the data and appears from the very definition of reuse (Bassi, 2011, p. 67).

In light of the analysis of the data altruism mechanism, the contribution illustrates the legal challenges, which are: (1) the risk of fragmentation; (2) the security concerns; (3) the duty of control on data altruism organisations.

(1) The risk of fragmentation is generated by the envisaged difficulties in implementing data altruism. Even though the DGA is a Regulation, in the implementation of data altruism the Member States play a decisive role. While waiting to understand how they will implement the data altruism mechanism, it is worth analysing the possible Member States' strategies and approaches.

(2) Concerning the security, the data altruism organisation must also ensure a solid infrastructure system. The goal is to create pools of data and this data must be stored, transferred, and managed, which makes the infrastructure absolutely central. The centralisation of data always brings with it several challenges from a security viewpoint, making those holding the data both very powerful, and at the same time very vulnerable. Very powerful, because it generates "the emergence of pools of data made available on the basis of data altruism that have a sufficient size in order to enable data analytics and machine learning, including across borders in the Union" (Recital 45, DGA). Highly weak because they are more easily targeted by cyber-attacks and data breaches.

(3) Then, the Article 21 of the DGA establishes a duty to control in charge of any data altruism organisations, over the third parties that are allowed to process the data. Although this requirement is understandable on principle, it does not seem easy achievable in practice. This mechanism of mutual controls seems to strongly refer to the mechanisms of accountability of the GDPR which establish, for the data controller, a set of duties, also with regard to the data processors, those who actually process the data, in the name and on behalf of the data controller (Durante, 2021, p. 134). Admittedly, data altruism organisations must only ensure that these users conduct processing for purposes of general interest. However, the DGA itself considers such purposes in a very broad and general way, referring to scientific research or the

[12] 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 (General Data Protection Regulation) (Text with EEA relevance) ELI: http://data.europa.eu/eli/reg/2016/679/oj.

improvement of public services. These categories are very broad, and without the identification of further boundaries, much can fall within the all-encompassing terms set by the European Regulation.

From the ethical perspective, the contribution intends to shed light on three aspects: (1) the terminological uncertainty; (2) the decoding the concept of altruism; and (3) the individual autonomy of the consent.

(1) As emerged from the description of the different phases of the data altruism mechanism according to the DGA, the concept of 'general interest' is crucial. However, there is terminological uncertainty insofar as the concept of 'public interest' emerges in the proposal of the Regulation. The 'public interest' diverges from the concept of 'general interest' and entails considerable debate if conceived according to the GDPR. In addition, the explanatory memorandum also mentions the concept of 'common good', amplifying the terminological uncertainty that impacts the ethics underlying the data altruism mechanism.

(2) Furthermore, the paper intends to develop the analysis regarding the application of the notion of altruism to the release of data, referring to the studies conducted on the so-called 'data philanthropy' (Taddeo, 2016; Taddeo 2017; Giannopoulou, 2019) that pave the way for the assessment regarding the practical feasibility of this mechanism (Veil, 2022). In light of a well-established trend, it is not difficult to envisage that there might be a fair amount of success on the side of data subjects and data holders who release their personal and non-personal data for the pursuit of general interest purposes (Ienca, 2023, p. 2). Several experiences show a general inclination to release more easily personal data for scientific research purposes (Pagallo, 2022, p. 74). Similarly, it is not difficult to envisage data users, which may include universities, research centres, but also private companies, foundations, etc. It is more difficult to identify entities undergoing the registration process to become data altruism organisations. For this reason, the role of potential private actors performing the function of data altruism organisations and the impact, from an ethical perspective, on the integrity of scientific research deserve further investigation.

(3) Finally, the data altruism is a consent-based mechanism. The goals of the introduction of the GDPR, compared to the previous discipline of the Directive 95/46/EC, was precisely to overcome the model of personal data processing primarily based on consent. The Article 6 of the GDPR provides for a set of mandatory legal bases for the processing of personal data: consent thus becomes one of the possible bases. The *ratio* for this choice made by the European lawmaker in 2016 was precisely to replace a consent-based approach that had proved to be ineffective (Solove 2012; Schermer, *et al*. 2014). It will therefore be necessary to further investigate the impact of the use of consent at the basis of the data altruism mechanism in relation to the choice of the legal bases required for processing personal data for scientific research purposes, according to the GDPR. The analysis of the role of consent in the altruism of data is relevant to the extent that it generates an impact on the personal autonomy of the individual giving such consent.

Given the set of challenges, both legal and ethical, and the multiplicity of actors involved in the data altruism mechanism, it is worth investigating how and whether to develop governance mechanisms that are able to hold together all the aspects at stake.

KEYWORDS: Data governance act, DGA, data governance, data altruism, scientific research, public interest.

REFERENCES

Bassi, E. (2011). PSI, protezione dei dati personali, anonimizzazione. *Informatica e diritto* 37.1-2, 65-83.

Durante, M. (2021) *Computational power: the impact of ICT on law, society and knowledge*. New York: Routledge.

Giannopoulou, A. (2019). Access and Reuse of Machine-Generated Data for Scientific Research. *Erasmus Law Review* 2, 155-165.

Ruohonen, J., Mickelsson, S. (2023). Reflections on the Data Governance Act. *Digital Society* 2.1, 1-10.

Ienca, M. (2023) "Medical data sharing and privacy: a false dichotomy?", *Swiss Medical Weekly* 153.1, 1-3.

Pagallo, U., Bassi, E. (2013). Open Data Protection: Challenges, Perspectives, and Tools for the Reuse of PSI, in Hildebrandt, M., *et al*. (eds), *Digital Enlightenment Yearbook 2013.* Amsterdam: IOS Press, 179-189.

Pagallo, U. (2020). The Politics of Data in EU Law: Will It Succeed? *Digital Society* 1.3, 1-20.

Pagallo, U. (2022). *Il dovere alla salute. Sul rischio di sottoutilizzo dell'intelligenza artificiale in ambito sanitario*, Milano-Udine: Mimesis.

Schermer, B. W., *et al*. (2014) The crisis of consent: How stronger legal protection may lead to weaker consent in data protection. *Ethics and Information Technology* 16.2, 171-182.

Solove, D. J. (2012) Introduction: Privacy self-management and the consent dilemma. *Harvard Law Review* 126, 1880-1903.

Taddeo, M. (2017). Data philanthropy and individual rights. *Minds and Machines* 27.1, 1-5.

Taddeo, M. (2016). Data philanthropy and the design of the infraethics for information societies. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2083, 1-12.

Van Eechoud, M. (2021). A Serpent Eating Its Tail: The Database Directive Meets the Open Data Directive. I*IC - International Review of Intellectual Property and Competition Law,* 52, 375-378.

Veil, W. (2022). Data Altruism: How the EU is Screwing up a Good Idea. Discussion paper. AlgorithmWatch, 1-8.

LOOKS LIKE A HUMAN, ACTS LIKE A HUMAN, BUT IS IT SOMETHING ELSE? AI AS SCHEIN-DASEIN

Jani Koskinen, Salla Westerstrand and Juhani Naskali

Information Systems Science, University of Turku, Turku (Finland)

jasiko@utu.fi; salla.k.westerstrand@utu.fi; juhani.naskali@utu.fi

EXTENDED ABSTRACT

Ethics of systems utilising Artificial Intelligence (AI) is an increasingly discussed topic in academia (e.g., Franzke, 2022), industry (e.g., Morley et al., 2021; Jobin et al., 2019), and popular media (Ouchchy et al., 2020). The recent popularisation of AI, following the introduction of new solutions with easier user interfaces, such as ChatGPT, has further accelerated the discussion. How do these technologies influence people's lives? Can AI have moral agency? How should we interact with the technology when it reminds us of our fellow humans? Can we soon evem make such distinctions, and if not, what could that mean for our moral agency?

To shed light on these complex questions, one option is to lean on Martin Heidegger's work and engage in an ontological discussion on the role of AI systems through the concept of *Dasein*.

Dasein is the central term for Heidegger (1927). He discussed the concept of being under deep and permanent ontological investigation, and used it to describe human existence that has awareness and confronts its own being in this world. Heidegger presents three modes of being in *Being and Time*, namely ready-to-hand, present-at-hand and Dasein, which all differ from each other with their unique characteristics. He did not offer strict or explicit explanations to being in *Being and Time* because the project was never entirely completed. Instead, he attempted to bring clarity to the question from different perspectives, emphasising the individual comprehension: for Heidegger being is based on hermeneutical phenomenology and—in simple terms—this means that being can only be investigated from the first-person perceptive. Only people themselves can reach an understanding of their Dasein. (see Heidegger, 1927)

For an ontological analysis of AI, it is essential to understand the three primary modes of being defined by Heidegger. Doing so reveals that AI systems could be giving a rise to something novel: a fourth mode of being.

Heidegger describes things (objects) and their being by a hammer example. First, Heidegger explained that something is *ready-to-hand* if it has some purpose to accomplish – like a hammer is used for hammering (Heidegger, 1927, §15–18). Usually, we do not give much consideration to the objects we use; we just use them like we always have and accept that they are there, ready for us to use to accomplish a certain goal. For example, when you are reading an article, the tool (the paper or the screen) that allows you to read it is not used consciously. You just use it and hopefully concentrate on the content of the article and get some sense out of it (the goal, or the purpose). Thus, we use such objects in the way they are meant to be used—or should we say, how they are properly used.

The second mode of being – *present-at-hand* – can be exposed by the breaking of an object. Brokenness reveals the object and exposes its natures, which refers to the purpose for which

the thing exists (see Heidegger, 1927, § 16). The term referral indicates that we understand the meaning of an object by its reference: for example, a hammer is referring to nails and wood towards the wall under construction. When the hammer breaks, we become conscious of its nature – it is revealed for us. When the hammer is not broken, we do not give it much thought, and it is revealed as ready-to-hand. Heidegger (1927, §18) shows that objects that are ready-to-hand appear to the observer in the context of the surrounding world and are referred to, along with other things, in the world for some purpose. Entities have significance only in their full context: a knife is a different thing in the kitchen, in a theatre, or in the hands of a criminal (Harman, 2010).

What makes situating an AI system – such as transformer-based language models like ChatGPT – in the hammer example difficult is that the being of AI does not seem to limit to the ready-to-hand. Instead, AI systems are something that to a human eye resembles Dasein, or 'the individual human mode of being in the world', which is one of the ways to grasp and present the meaning of the original German term (). The special character of Dasein compared to other two modes of being is that Dasein is the only one that can have an understanding of one's own being and hence can also investigate it. Thus, Dasein is a mode of being that is traditionally associated with (only) human beings (Van Der Hoorn and Whitty, 2015). This understanding of one's existence is the key factor that separates Dasein from present-at-hand and especially from ready-to-hand. Dasein can see the present-at-hand and the ready- to-hand, but Dasein cannot truly be reached as present-at-hand or as ready-to-hand. Things, or artefacts can be present or ready but only Dasein (human) can see other modes and give meanings for those.

Hence, we argue that AI has given birth to a fourth mode of Being: *ScheinDasein* (looking-like-Dasein), that reveals itself in such ways that it seems like Dasein – a human behind the technology. They may appear as witty conversationalists, therapists, or even romantic partners (Hale, 2023; Cost and Court, 2023). AI can even seem to be able conduct deep self-investigations (a key factor separating people as Dasein form objects) that ordinary people cannot easily achieve because of our human limitations. The interaction with AI can, at its best (or worst, depending on the situation), give people new insights and provoke feelings of empathy and of being understood.

In the future, it is possible that people will not be able to distinguish between actual Dasein and Schein-Dasein—although the idea of seeing Dasein is a paradox in itself, as Dasein is always a lived experience by individuals, by themselves. As a consequence, due to the ongoing popularisation of ever more pervasive AI systems (), we may end up in a situation where our being (Dasein) is left alone as we cannot be entirely sure that we are living with other selfconscious people. Instead, we may feel like we are left alone with mere objects. This also set problem with death as possibility for us as Dasein.

Like Heidegger (Heidegger, 1927, §51–53) shows us, death is something that only Dasein can and must face, and it should not be confronted like the ordinary man (*das Man*) does. Das Man is a term that Heidegger (1927) uses to describe a situation where people consciously choose to hide or lose themselves and replace themselves with commonly given ways of being or acting, whereas Dasein is living a life consciously and actively makes sense of it. Thus, das Man could be described as a generally accepted and non-disturbing way of living or being. However, death is an issue which cannot be outsourced to das Man, because common shared way of living cannot reach or face the death. Actually, das Man gives justification and adds temptation to cover up oneself from one's own most possibility as being-towards-death (*Sein-zum-Tode*)

(Heidegger, 1927). By being-towards-death one could see what is important and how one wants to spend life, for example, with family.

Yet, in the case of AI, Death has a very different meaning. We are not sure anymore if we are living with people or Shein-Dasein—objects that deceive us to make false conclusions of our surroundings. This could lead to a Dasein turning into das Man, who is not able to reflect consciously and make sense of itself or its surroundings. We merely believe that we are Dasein in this life and have, for example, decided to be with our family (which could turn out to be a collection of AI systems) because it was what seen worthwhile as a beingtowards-death who recognises that having a limited lifespan worth spending wisely. Furthermore, AI makes it possible to claim Shein-Dasein as our own– to ”create” art by tasking an AI to do it, or ”write” a book requesting it from an AI, claiming the apparent creativity as our own, living as das Man with the outer appearance and self-esteem of Dasein.

Introducing human-like AI systems raises fundamental ethical questions: if we can no longer distinguish Schein-Dasein from Dasein, how much of human autonomy do we have left? Can we ever make rational decisions and interact with others in a way that enables ethical action, which would be a requirement of, e.g., in Habermasian discourse ethics? Do we have moral obligations towards Schein-Dasein, like we would have towards Dasein? When does Schein-Dasein, in fact, become Dasein – if ever? Such a fourth mode of being requires further examination.

KEYWORDS: AI, Dasein, Heidegger, Ontology.

REFERENCES

Cost, B. & Court, A. (2023). My girlfriend was really an ai catfish – i feel cheated. https://nypost.com/2023/04/12/my-girlfriend-was-really-an-aicatfish-i-feel-cheated/

Franzke, A. S. (2022). An exploratory qualitative analysis of ai ethics guidelines. *Journal of Information, Communication and Ethics in Society*. https://doi.org/10.1108/JICES-12-2020-0125

Hale, E. (2023). Chatgpt is giving therapy. a mental health revolution may be next — technology — al jazeera. https://www.aljazeera.com/economy/2023/4/27/could-your-next-therapistbe-ai-tech-raises-hopes-concerns

Harman, G. (2010). Technology, objects and things in heidegger. *Cambridge journal of economics*, 34(1):17–25. https://doi.org/10.1093/cje/bep021

Heidegger, M. (1927). *Sein und Zeit*. Used several translations. Main translation *Oleminen ja Aik*a by Kupiainen R. 2000. Tampere: Vastapaino.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399. https://doi.org/10.1038/s42256-019-0088-2

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2021). Operationalising ai ethics: barriers, enablers and next steps. *AI & Society,* 38, 411–423. https://doi.org/10.1007/s00146-021-01308-8

Ouchchy, L., Coin, A., and Dubljevi´c, V. (2020). Ai in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY*, 35: 927–936. https://doi.org/10.1007/s00146-020-00965-5

Van Der Hoorn, B. & Whitty, S. J. (2015). A heideggerian paradigm for project management: Breaking free of the disciplinary matrix and its cartesian ontology. *International Journal of Project Management*, 33(4):721–734. https://doi.org/10.1016/j.ijproman.2014.09.007

PRIVACY AFTER *DOBBS*: HOW THE SHIFTING U.S. LANDSCAPE AFFECTS THE BROADER DEBATE

Michael S. Kirkpatrick

James Madison University (USA)

kirkpams@jmu.edu

EXTENDED ABSTRACT

In June 2022, the Supreme Court of the United States ("SCOTUS") released its decision in *Dobbs v. Jackson Women's Health*, overturning two earlier decisions (*Roe v. Wade* in 1973 and *Planned Parenthood v. Casey* in 1992. The most immediate focus of these three decisions centers around legal protections for abortion throughout the U.S. Under *Roe*, no state (or the federal government itself) could pass a law that restricted abortion in the first two trimesters of pregnancy. The *Casey* decision mostly upheld *Roe*, although it allowed states to pass certain restrictions so long as they did not pose an "undue burden" on pregnant women. These two decisions established the right to abortion as having a foundation in the U.S. Constitution that could not be undermined through basic legislative action. The *Dobbs* decision overturned both *Roe* and *Casey*, thereby declaring these earlier decisions invalid. Abortion was no longer considered to be a fundamental right protected by the Constitution, allowing states to pass laws that would completely ban abortion, which many states did.

Although these three decisions and their related debates are primarily focused on the legality of abortion, they are more properly understood as decisions about the nature of privacy and whether privacy is considered to be a fundamental right in the U.S. In the case of *Roe* and *Casey*, the protection of abortion was an indirect effect of an implicit right to privacy. In both decisions, the right to privacy was determined to be implied by the 14th Amendment's protection of the right to due process. Specifically, according to the legal doctrine of *substantive* due process (as opposed to *procedural* due process), states and the federal government are restricted from passing laws that arbitrarily intrude into citizens' private lives. Both decisions also relied on an earlier decision, *Griswold v. Connecticut*, that described the right to privacy as being part of the *penumbras* of certain explicitly mentioned rights rather than the due process clause. In essence, the interpretation of due process in the *Roe* and *Casey* decisions, as well as the penumbras described by *Griswold*, is comparable to Article 8 (Right to respect for private and family life) of the European Convention on Human Rights. The *Dobbs* decision explicitly argued that the idea that the right to privacy is inherent in the protections of the due process clause is "egregiously wrong" and therefore both decisions must be overturned. (*Dobbs* did not overturn *Griswold*.)

One central point of divergence between the *Dobbs* decision and the findings of *Roe* and *Casey* is a disagreement about the nature of privacy itself. According to the author of the *Dobbs* decision, *Roe* relied on a philosophical view of privacy that "conflated the right to shield information from disclosure and the right to make and implement important personal decisions without governmental interference." In other words, the debate rests on whether "privacy" is fundamentally about *secrecy* or *autonomy*. While the *Dobbs* authors adopted the former, which is a considerably narrower construction of privacy, the authors of *Roe* and *Casey*, along with many other scholars, adopted a broader view that leans toward the latter. For instance, Richards

(2022) identifies the distinction between multiple forms of privacy, including *spatial* privacy, *decisional* privacy, and *information* privacy. (Citron (2022) describes *intimate* privacy as another form.) The *Dobbs* authors essentially argued that decisional privacy does not have a constitutional basis.

This debate surrounding the nature of privacy has a long history, particularly in the U.S. where there is no explicit mention of privacy in the Constitution. One of the earliest and most well-known discussions is the characterization by Warren and Brandeis (1890) of privacy as the "right to be let alone," though others found the discussion to be vague and unhelpful. For instance, Thomson (1975) argued that "the most striking thing about the right to privacy is that nobody seems to have any very clear idea what it is." Thomson proposed addressing this lack of clarity by focusing on "a cluster of rights" that are primarily focused on "the right over the person" (i.e., the physical body) and rights concerning "owning property." In more recent work, Solove (2007) agreed that privacy was not a singular right but rather that it is "best used as a shorthand umbrella term for a related web of things," though Thomson's limited focus on the person and property miss many important privacy invasions.

In contrast, Rachels (1975) emphasized that "there is a close connection between our ability to control who has access to us and information about us, and our ability to create and maintain different sorts of social relationships." In other words, the purpose of controlling information (secrecy) is about freely interacting with society (autonomy). Nissenbaum (2010) more explicitly links privacy with autonomy, as limiting access to information "contributes to material conditions for the development and exercise of autonomy and freedom in thought and action." Citron (2022) argues that intimate privacy "is a precondition to a life of meaning."

In short, the *Dobbs* decision marks a turning point in U.S. law regarding privacy. The trend in both scholarship and law had been toward broadening the concept of privacy toward a more expansive right beyond simply information privacy, and *Dobbs* has stopped that trend. Although the immediate effect is on the legality of abortion in the U.S., other SCOTUS decisions also relied on the foundation of privacy in the due process clause. As such, it is not clear at this point to what extent this decision will affect the privacy debate.

It should be noted that, although this history is focused on the U.S. perspective, the full impact of the *Dobbs* decision will be international. Many people have noted that this decision will shape how technology companies implement and maintain privacy (Federman, 2022; Krishnan et al., 2022; Privacy International, 2022; Sexton, 2022), how medical organizations protection patient information (Clayton et al., 2023; Henneberg, 2022), and how information gathered from technology companies will affect law procedures (Edelson, 2022; Kamin, 2023; Marathe, 2022; Stuart, 2023). The Internet is a global network, so the capabilities that technology companies build in the U.S. will impact the services and protections that they can provide in other parts of the world.

In this talk, we will discuss the evolution of the concept of privacy through legal scholarship, focusing on how *Dobbs* influences that debate. We will also discuss the multiple forms of privacy (including decisional privacy) and how the U.S. and E.U. differ in their approaches. Finally, we will highlight concerns about how the shifting U.S. legal approach may impact privacy protections on the Internet moving forward.

KEYWORDS: Privacy, decisional privacy, SCOTUS, substantive due process.

REFERENCES

Citron, D. K. (2022). *The Fight for Privacy*. Norton Books.

Clayton, E. W., Embi, P. J., & Malin, B. A. (2022). Dobbs and the future of health data privacy for patients and healthcare organizations. *Journal of the American Medical Informatics Association, 30(1)*, 155–160. https://doi.org/10.1093/jamia/ocac155

Edelson, J. (2022, September 22). Post-Dobbs, your private data will be used against you. *Bloomberg Law.* https://news.bloomberglaw.com/us-law-week/post-dobbs-your-private-data-will-be-used-against-you

Federman, H. (2022, September 29). Privacy and data protection in the wake of Dobbs. *Security*. https://www.securitymagazine.com/articles/98414-privacy-and-data-protection-in-the-wake-of-dobbs

Henneberg, C. (2023, June 5). The trade-offs for privacy in a post-Dobbs era. *Wired*. https://www.wired.com/story/the-trade-offs-for-privacy-in-a-post-dobbs-era/

Joh, E. E. (September 5, 2022). Dobbs online: Digital rights as abortion rights. In (Levendowski, A. & Jones, M. L. (eds.), *Feminist Cyberlaw,* forthcoming 2023. Available at SSRN: https://ssrn.com/abstract=4210754 or http://doi.org/10.2139/ssrn.4210754

Kamin, S. (2022, December 18). Katz and Dobbs: Imagining the Fourth Amendment without a right to privacy. *Texas Law Review*. https://texaslawreview.org/katz-and-dobbs-imagining-the-fourth-amendment-without-a-right-to-privacy/

Krishnan, A., Cohen, K., & Hackley, C. (2022, August 27). Digital privacy in the post-Dobbs. *The Regulatory Review*. https://www.theregreview.org/2022/08/27/saturday-seminar-digital-privacy-in-the-post-dobbs-landscape/

Marathe, I. (2022, July 1). Post-'Dobbs,' privacy attorneys prepare for increased data surveillance *Legaltech News*. https://www.law.com/legaltechnews/2022/06/27/post-dobbs-privacy-attorneys-prepare-for-increased-data-surveillance/

Nissenbaum, H. (2010). *Privacy in Context*. Stanford Law Books.

Privacy International. (2022). Privacy and the body: Privacy International's response to the U.S. Supreme Court's attack on reproductive rights. *Privacy International*. https://privacyinternational.org/news-analysis/4938/privacy-and-body-privacy-internationals-response-us-supreme-courts-attack

Rachels, J. (1975). Why privacy is important. *Philosophy & Public Affairs, 4(4)*, 323–333. http://www.jstor.org/stable/2265077

Richards, N. (2022). *Why Privacy Matters*. Oxford Books.

Sexton, M. (2023, January 22). The new front in the battle for digital privacy post-Dobbs. *Third Way*. https://www.thirdway.org/memo/the-new-front-in-the-battle-for-digital-privacy-post-dobbs

Stuart, A. H. (October 26, 2022). Privacy in discovery After Dobbs. *Virginia Journal of Law and Technology*. http://doi.org/10.2139/ssrn.4259508

Thomson, J. J. (1975). The right to privacy. *Philosophy & Public Affairs, 4(4)*, 295–314. http://www.jstor.org/stable/2265075

Warren, S. D. & Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review, 4(5)*, 193–220. https://doi.org/10.2307/1321160

TOWARDS AN AIMLESS EXISTENCE – A DIALOGUE ABOUT AI'S POTENTIAL TO RADICALLY CHANGE THE HUMAN CONDITION

Mikael Laaksoharju, Iordanis Kavathatzopoulos

Department of Information Technology, Uppsala University (Sweden)

Mikael.Laaksoharju@it.uu.se; Iordanis.Kavathatzopoulos@it.uu.se

EXTENDED ABSTRACT

This presentation will be given in the form of a dialogue between Laaksoharju and Kavathatzopoulos. Let us start with an introduction.

In the recent years, dystopian prophecies regarding artificial intelligence (AI) have garnered public attention. For instance, the risk of AI becoming exponentially more powerful than all human intelligence combined, acquiring an independent existence of itself, transforming us into something we do not want to be, evolving in a radically different way, even affect the whole universe, etc. (Kurzweil, 2006; Bostrom, 2014; O'Neil, 2016; Harari, 2016; Reese, 2018; Tegmark, 2017; see also Future of Life Institute, 2023). Although not everyone agrees on whether any of these things will happen, or when they might happen, these prophets have in common that they focus mainly on the technical aspects of the issue or on AI itself and its purported potential. There is, however, another interesting – and in our view more relevant – angle to AI as a phenomenon, namely the effects that even weak but well-functioning AI could have on human nature, or life in general.

The complexity and opacity of many AI algorithms is often called out as a great risk of potentially losing control over the algorithms. Consequently, large research efforts have been invested in what is called "Explainable AI". We do not wish to dispute the importance of this research area but perhaps additional considerations can be added to nuance the ambition.

First of all, complexity and opacity in themselves do not necessarily imply loss of control, if the algorithms are completely predictable. For instance, most people do not know how the technology in a regular car works and the trend has been that in every new generation of cars even more of the underlying technology is hidden from car owners. As long as the car functions as expected, this is likely net beneficial for the cognitively overloaded (post)modern individual.

Some will argue that the trend of hiding technological complexity threatens human autonomy and indeed there is a sacrifice of autonomy in, e.g., giving up some control over your vehicle. Nevertheless, most seem to consider that the benefits outweigh the minor harm in giving up low-level control. Judging from the public attitude, this will be a likely fate of algorithms as well when they start producing consistently reliable results. Most of us will have no interest in scrutinizing the process behind an algorithm's recommendation or classification when it is perceived as accurate.

However, this means a cognitive cost. When algorithms will be perceived as reliable, they will start entering *the fabric of truth production*, much like calculators have been elevated to determining the correct result of arithmetic calculations and how statistical tests have come to represent the existence of correlations. The difference here is that AI algorithms can tell the "truth" about so much more than statistical tests. They will be able to decide for us whether we

are looking at a picture of a sloth or a chocolate croissant (see e.g. Zack 2016, Alasadi 2019). One day we may have become so used to trusting the algorithm so that instead of musing over how similar some sloths and some chocolate croissants look in some photos, we will be fascinated with how some photos of croissants actually could have been photos of sloths, and vice versa, if the computer did not tell us the truth. When algorithms are accurate almost every time, we will quickly lose our current skepticism towards them, simply because skepticism is unnecessarily burdensome. In other words, what the algorithms tell us is the truth will be the most convenient belief. Solomon Asch (1956) would not become surprised.

Here it is time to introduce the different positions of Laaksoharju and Kavathatzopoulos. Kavathatzopoulos (2024) claims that potent (weak) AI could become an existential threat by fulfilling our needs to the extent that the human capability to reason will eventually wane. After all, if any human goal can be fulfilled by AI, why would we ever need to practice our reasoning ability? In a sense, this is a philosophically and psychologically founded Wall-E prophecy of the future.

Laaksoharju claims that this prediction is based on an overly teleological assumption of both human behaviour and of AI. When it comes to humans, goal fulfillment, as a construct, is not a necessary condition for activating thinking, and when it comes to AI, the goals that are currently formulated and assessed are in the form of tasks for which there exist ground truths that have been somewhat arbitrarily decided by humans. The implication of this is that the current success of algorithms is more of a social construct than something that corresponds to any actual human need; a self-selected group of arbiters' have chosen what problems are relevant to be processed by AI and then deemed these problems as solved to some extent.

Before proceeding, it should be mentioned that the positioning of human nature outside of, or in opposition to, technology is limiting. In line with the views of Bernard Stiegler (1998), we see technology as a response to human/organizational/societal values and by that entangled with human nature. Humans do not primarily interact with technology but their interaction with other humans is augmented/mediated/supplemented by technology. With this lens, nuclear weapons, for instance, are arguments in negotiations about territorial power, and AI algorithms are arguments in negotiations about power in general.

In essence, the dialogue is revolving around the concept of goals and its importance for human existence. If Kavathatzopoulos is correct, even weak but effective AI will lead to the demise of humanity. If Laaksoharju is correct, strong AI is still a pipe dream and weak AI will be just like any technology introduction – it will change the logic of existence to some extent but humans will find new ways to compete with each other. The questions to be addressed in the dialogue are thus:

1. Do goals "exist" or not?
   Kavathatzopoulos claims that goals are an integral part of thinking/life, which emerged because of uncertainties in the kinesis of the world. However, if they do exist, they are either real (which conflicts with the perception of the world as chaotic motion, which itself is a prerequisite for the existence of goals since goals have meaning in a world of uncertainty; real goals seem to be a contradiction in terms) or the goals are an illusion of thinking (which is in accordance with the world as chaos, i.e., goals arise in uncertainty, as something to be sought, identified, and pursued, but they cannot be real because then the whole process/thinking would be "locked").

Laaksoharju simply refutes the explanatory value of “goals” for understanding human behavior and instead claims that humans largely act by following mental patterns that are activated by stimuli in their lifeworlds. However, Kavathatzopoulos means that there is no explanatory value of “goals” but they are there together with the thinking process, which is not possible to run without both of them.

2. Are self-determined goals sufficient for AI to become "autonomous"?
   Kavathatzopoulos will argue for the possibility that if goals arise in connection with uncertainty and life emerges, perhaps facilitating "goals" for AI will open up the possibility for AI to have its own "life". Perhaps the more one "confuses" AI regarding which goals to strive for or invent on its own, the harder it will be for AI to become autonomous.

   Laaksoharju will argue that goals can be useful for programming sensing systems to determine how to regulate their behaviors, but that this goal-directed behavior will not lead to anything similar to human consciousness in machines.

   The ironical conclusion of this introduction is that the predictions of both Laaksoharju and Kavathatzopoulos will lead to an understanding of human existence as aimless, with the difference that we are either experiencing it already now or we will as soon as we have perfected AI to fulfill all our desires.

KEYWORDS: Artificial general intelligence, motivation, existential threat, alignment problem.

REFERENCES

Alasadi, Z. (2019). *Is it a Sloth or a Chocolate Croissant*? Github. https://github.com/zainabalasadi/sloth-or-croissant

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and applied*, 70(9), 1.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Future of Life Institute. (2023). *Pause giant AI experiments: An open letter.* https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Harari, Y. N. (2016). *Homo Deus: A Brief History of Tomorrow.* Random House.

Kavathatzopoulos, I. (2024). 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). London: Routledge.

Kurzweil, R. (2006). *The Singularity is near: When humans transcend biology.* Penguin Books.

O’Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown.

Reese, B. (2018). *The fourth age: Smart robots, conscious computers, and the future of humanity.* Atria Books.

Stiegler, B., 1998, *Technics and Time, 1: The Fault of Epimetheus*, Stanford: Stanford University Press.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence.* Knopf Publishing Group.

Zack, K. (2016, March 10). *Chihuahua or muffin* [Tweet]. Twitter. https://twitter.com/teenybiscuit/status/707727863571582978

PREDICTING THE UNPREDICTABLE- THE ETHICS OF DIGITAL FINANCE

Gonçalo Costa

Aroska (Portugal)

goncalo.costa@aroska.pt

EXTENDED ABSTRACT

Introduction

Digital technologies underway are reshaping societies today and in the future. Economy digitalisation is an ongoing and continuous process, which promises to spur innovation, generate efficiencies, and improve services. Besides, a successful transition is a key condition to boost more inclusive and sustainable growth; although, digitalisation can be a disruptive process with unforeseen results (OECD, 2018).

Digital finance is a blurred and multidimensional concept under debate which the following data demonstrate:

a) "annual data generation is estimated to be doubling every year, and the overall size will reach 44 zettabytes (that's trillions of gigabytes) by 2020" (Bayat-Renoux, Svensson & Chebly, 2021), as well as it is expected the world generate 181 zettabytes by 2025 (Pointing, 2022);

b) digital finance potential boost of emerging economies` GDP (Gross Domestic Product) expectation is $3.7 trillion by 2025, with a six per cent increase versus a business-as-usual scenario (Khera, 2021);

c) Artificial Intelligence (AI) alone could lift global GDP by an estimated US$15-20 trillion by 2030 through digital finance optimisation (Report Linker, 2023).

However, some doubts arise: i) digital finance will be a novel concept?; ii) which data or statistics acknowledge it?; iii) to what extent digital finance will occur?; iv) which potential technologies fuel it?; and v) potential ethical issues?

Digital finance

Definition

In the literature, there is no agreement on the digital finance definition despite a wide spectrum of international and national institutions, as well as private companies that lead with a large range of novel products or technologies in finance. Some keen examples are:

a) digital financial services are an ecosystem consisting of users who require digital and interoperable financial products and services through digital means from providers. These providers are responsible for financial, technical and other infrastructures; while, complying with laws and regulations to make such services accessible, affordable, and safe (ITU, 2023);

b) digital financial services can integrate any financial process through digital technology, which can include electronic money, mobile or online financial services, iteller solutions, and branchless banking (European Union, 2023);

c) digital finance acknowledges services delivered over digital infrastructure with low usage of cash and traditional bank branches. Digital infrastructure encompasses whole devices that connect individuals and businesses to digitized national payment infrastructure, facilitating transactions across all parties (Feyen *et al.*, 2021).

Therefore, the author will attempt to draw conceptual boundaries in interconnected definitions. Some examples are “e-money”, “electronic banking”, “mobile banking”, “fintech”, etc.

Conceptual boundaries

To draw conceptual boundaries several procedures are required: i) define each related concept; ii) understand their potential relationships; and, iii) the author rationale.

E-money is as an electronic store of monetary value for making payments to entities other than the e-money issuer (European Central Bank, 2023). E-banking is a procedure between a bank/financial institution and its customers for encrypted transactions through the web or, customer basic requirements (personal data, balance inquiry or account state) (Team FinFirst, 2022). Mobile banking is widely recognised as e-banking services through APPs in order to retort customers demand (mobility and immediate access) (CFI Education, 2021). Fintech refers to a myriad of technologies to augment, streamline, digitize or disrupt traditional financial services. Some examples are: i) make a deposit through a snapshot of a paycheck; ii) peer-to-peer lending; or, iii) immediate currency exchange (Walden, 2022).

This interconnected *continuum* is strongly aligned with the ITU definition; however, it encompasses numerous social and ethical dilemmas (equity, digital divide, security, etc).

What to predict?

Predict is guessing! Although, despite the shade, it is possible to shed some light upon forthcoming technologies in financial services. IBM argues that cloud computing is becoming mainstream in banking, namely, to search the optimal mix between traditional IT, public and private clouds. “With hybrid cloud, banks have the flexibility and benefits of both private and public cloud, while addressing data security, governance, and compliance” (Marous, 2018a).

API platforms are changing entirely the banking ecosystem since financial institutions serve as platform. I.e., other stakeholders build their own applications using the bank’s internal data; so, traditional commercial or retail banking will be under pressure (Marous, 2018b). This process will be enhanced with robotic process automation (RPA), because it simplifies compliance by retaining detailed logs of automated processes, automatic reports to auditors or managers. Recent estimates denote that intelligent automation, a blend between machine learning and data patterns analysis, will reduce administrative and regulatory processes costs by at least 50% (Rajan, 2018; Donelly, 2022).

Instant payments technology is already available in some countries through P2P services, which are a tempting opportunity to achieve speed, experience, and availability that fulfil generation

Y consumers' expectations (Marous, 2018c). However, extended mobile solutions require a different digital structure, as well as, organizational in which Artificial Intelligence (AI) will play a decisive role. From the mashup explosive growth of structured and unstructured data, novel technologies (e.g., cloud computing, machine learning, etc.) several pressures arise (European Union, 2023).

Those pressures along with security and privacy enable blockchain technology; although, despite the transformational impact on the baking industry that some experts argue (Quindazzi, 2017) some recent cases regarding blockchain and cryptocurrencies deny it (Xu *et al.*, 2022). Therefore, a recent trend on cyber risks is prescriptive security, which explores AI and other tools to monitor, detect and stop in real-time potential threats before they strike (Streeter, 2018).

Marous (2017) also suggests that augmented reality (AR) and virtual reality (VR) can help bank customers autonomy in physical investments, i.e.., during a visit to a house, store or land immediate information on property sales, price tendencies, current listings, and properties selling or sold in the area is delivered. The scope is narrowed and most likely will occur for other financial products. Interconnected technologies to AR and VR such as smart vision systems, virtual assistants, natural language processing technologies will arise shortly. One example is Amazon's Alexa, a virtual assistant, for the Bank of America. Therefore, smart machines attempt to digitally engage customers through guidance and support (avoid customer loss) (Rock Paper, 2022).

At last, but not least, quantum computing will support the entire network or infrastructure. These represent a major leap forward in computing power, surpassing cloud or blockchain potential. JP Morgan and Barclays have an agreement to investigate quantum computing potential (Brown, 2016).

Gray predictions

Some institutions, for instance, the World Bank describe "digital finance" as an important milestone for societal inclusion; although also is a potential minefield regarding supervisory and regulatory actions (ITU, 2023), because digital finance is a sociotechnological-driven process. I.e.., requires education upon decision making, financial and technological literacy. The author believes that neuroeconomics will play a decisive role, since it is an interdisciplinary research field that explains decision making, multiple alternative procedures and what actions are to be followed (Rebecca & Belden, 2011)

KEYWORDS: Ethics, digital finance, digital technologies.

REFERENCES

Bayat-Renoux, F., Svensson, U. & Chebly, J. (2021). *Digital technologies for mobilizing sustainable finance- Applications of digital technologies to sustainable finance*. G20 Sustainable Finance Study Group (SFSG). Mava Foundation.

Brown, J. L. (2016, July 28). Will quantum computing help government agencies improve cybersecurity? *Fed Tech Magazine*. Retrieved from https://fedtechmagazine.com/article/2016/07/will-quantum-computing-help-government-agencies-improve-cybersecurity

CFI Education (2023, September 30). Mobile banking- The use of a mobile device to carry out financial transactions. *CFI Education*. Retrieved from https://corporatefinanceinstitute.com/resources/wealth-management/mobile-banking/

Donnelly, S. (2022, September 15). 5 benefits of robotic process automation in finance. *Finance Alliance*. Retrieved from https://www.financealliance.io/5-benefits-of-robotic-process-automation-in-finance/

European Central Bank (2023, April 15). Electronic money. *European Central Bank*. Retrieved from https://www.ecb.europa.eu/stats/money_credit_banking/electronic_money/html/index.en.html

European Union (2023, September 30). Modernising payment services and opening financial services data: New opportunities for consumers and businesses. *European Union*. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/ip_23_3543

Feyen, E. *et al.* (2021, July 17). Fintech and the digital transformation of financial services: Implications for market structure and public policy. *Bank of International Settlements*. Retrieved from https://www.bis.org/publ/bppdf/bispap117.pdf

ITU (2023). *Digital financial services: Regulating for financial inclusion- An ICT perspective*. London: ITU.

Khera, P. *et al.* (2021, March 19). Digital finance inclusion in emerging and developing economies: A new index. *International Monetary Fund*. Retrieved from https://www.imf.org/en/Publications/WP/Issues/2021/03/19/Digital-Financial-Inclusion-in-Emerging-and-Developing-Economies-A-New-Index-50271

Marous, J. (2017, June 16). Will augmented and virtual reality replace the bank branch? *The Financial Brand*. Retrieved from https://thefinancialbrand.com/65828/ar-vr-voice-chatbot-bank-branch-replacement-trends/?internal-link

Marous, J. (2018a, November 13). Platforms, products, and channels transforming financial services. *The Financial Brand*. Retrieved from https://thefinancialbrand.com/77163/banking-platforms-channels-digitalization-open-api/?internal-link

Marous, J. (2018b, June 21). Financial institutions must explore open banking options today. *The Financial Brand*. Retrieved from https://thefinancialbrand.com/73135/open-banking-platform-trends/?internal-link

Marous, J. (2018c, January 16). Financial institutions must explore open banking options today. *The Financial Brand*. Retrieved from https://thefinancialbrand.com/69892/mobile-banking-payments-trends/?internal-link

Pointing, A. (2022, November 8). The dark side of data. *Medium*. Retrieved from https://medium.com/totalenergies-digital-factory/the-dark-side-of-data-46dea5740f9a

Quindazzi, M. (2017, May 17). What is blockchain… And why should I care? *The Financial Brand*. Retrieved from https://thefinancialbrand.com/65247/blockchain-bitcoin-banking-trends/?internal-link

Rajan, S. (2018). Robotics and cognitive automation will keep banks from drowning on data. *The Financial Brand*. Retrieved from https://thefinancialbrand.com/70188/robotic-cognitive-automation-banking-adoption-trends/?internal-link

Rebecca, S. & Belden, A. (2018). Neuroeconomics and neuromarketing: Practical applications and ethical concerns. *Journal of Mind Theory*, 0, 2, Retrieved from http://www.aslab.upm.es/documents/journals/JMT/Vol0-No2/JMT_0_2-NEU-BELDEN.pdf

Report Linker (2023). *Global fintech industry 2023-2027*. London: Report Linker.

Rock Paper (2022, December 7). Augmented reality in finance: The unlikely dynamic duo. *Rock Paper Reality*. Retrieved from https://rockpaperreality.com/insights/ar-use-cases/augmented-reality-in-finance-the-unlikely-dynamic-duo/

Streeter, B. (2018, August 6). What is blockchain… And why should I care? *The Financial Brand*. Retrieved from https://thefinancialbrand.com/74044/mobile-banking-features-digital-security/?internal-link

Team FinFirst (2018, June 19). *Is e-banking the same as internet banking? Know the difference*. IDC First Bank. Retrieved from https://www.idfcfirstbank.com/finfirst-blogs/finance/what-is-e-banking

Walden, S. (2022, July 15). *What is fintech*. Forbes Advisor. Retrieved from https://www.forbes.com/advisor/banking/what-is-fintech/

Xu, W. *et al.* (2022). Blockchain and digital finance. *Financial Innovation*, 8(97). Retrieved from https://jfin-swufe.springeropen.com/articles/10.1186/s40854-022-00420-y

AN ANALYSIS ON AI ETHICAL ASPECTS FROM A STAKEHOLDER'S PERSPECTIVE

Sofia Segkouli, Maria Tsourma, Pinelopi Troullinou, Paola Fratantoni, Dimitris Kyriazanos, Anastasios Drosou, Dimitrios Tzovaras

Information Technologies Institute, Centre for Research and Technology Hellas, Thermi, (Greece), Information Technologies Institute, Centre for Research and Technology Hellas, Thermi, (Greece), Trilateral Research (United Kingdom), Zanasi & Partners (Italy), National center of scientific research "Demokritos" (Greece), Information Technologies Institute, Centre for Research and Technology Hellas, Thermi, (Greece), Information Technologies Institute, Centre for Research and Technology Hellas, Thermi, (Greece)

sofia@iti.gr; mtsourma@iti.gr; pinelopi.troullinou@trilateralresearch.com; paola.fratantoni@zanasi-alessandro.eu; dkyri@iit.demokritos.gr; drosou@iti.gr; dimitrios.Tzovaras@iti.gr

EXTENDED ABSTRACT

Artificial Intelligence (AI) technologies rapid development and use stimulate vigorous discussion about their potential, in different contexts and uses (Carvalho et al, 2022). The application of Artificial Intelligence (AI) by various information and communication technology (ICT) professions provides several benefits (Eurostat, 2022). More specifically, varied viewpoints lead to more robust AI systems (Liu et al, 2022). When people from various backgrounds cooperate, they contribute unique ideas and experiences that can improve AI technology development and implementation. Second, the diversity of ICT experts means that AI systems are intended to appeal to a broad spectrum of users, supporting their diverse wants and preferences. As a result, AI applications become more inclusive and user-friendly (Meyer & Henke, 2023). However, there is a lack of research on how to provide AI technology in a more aligned way with social ideals and promote justice, transparency, accountability, and inclusion by including various viewpoints. In this study, the successful case study of [masked due to blind review], A European Positive Sum Approach towards AI tools in support of Law Enforcement and safeguarding privacy and fundamental rights, targeted at developing pathways on the basis of ethical views of ICT and academic professionals for AI design and development. To this direction, different background and expertise has been gathered in order to have a broader and more comprehensive understanding of issues and concerns related to the use of AI-based technologies in the security domain. Moreover, in order to conceptualize in depth ethical and security controversies, diverse societal sectors have been engaged in the light of taking into consideration all the voices and perspectives when developing ethical and legal oriented policies.

A controversial issue nowadays raises the question of whether the diversity of information and communication technology (ICT) professionals can raise ethical concerns. From an initial point of view, this inclusion may more effectively recognize and resolve biases and discriminatory behaviours in AI systems. Diversity gives a variety of viewpoints and experiences, which aids in the identification and mitigation of biases in AI systems. The danger of developing discriminatory or unjust AI algorithms that disproportionately affect particular persons or groups can be reduced by incorporating specialists from diverse backgrounds. Diverse viewpoints also improve

decision-making and encourage more inclusive AI systems that respond to a larger spectrum of users' requirements.

Furthermore, understanding the socioeconomic and cultural settings in which these technologies are employed is required for ethical issues in AI research. Diverse ICT experts contribute a plethora of cultural, social, and ethical understanding that may be used to inform AI system design and deployment. This guarantees that AI technology is consistent with local values, conventions, and legal frameworks, preventing ethical conflicts or harm. In this context, the use of AI by diverse professionals fosters transparency and accountability, because different viewpoints and expertise contribute to making AI systems explainable, auditable, and subject to critical examination.

On the other hand, the ethical views on the use of AI by diverse ICT professionals can vary based on individual perspectives and cultural backgrounds. In terms of [masked due to blind review] case study systematic surveys have been conducted to unlock specific ethical issues and concerns both at local level to identify methods and strategies of single countries but also compare different perceptions and feelings of similar topics. To this end LEAs (Law Enforcement Agencies) have been engaged along with relevant experts through policy labs.

In particular, academics and practitioners experience from different lenses and perspectives the potential implications of adopting AI-based technologies. One concern is the potential for biased outcomes in AI systems, as stated before. If professionals do not address biases in training data or fail to account for diverse perspectives during system development, AI can perpetuate existing societal biases and discrimination. Another concern is privacy and data protection. With diverse ICT professionals working on AI, there is a need to ensure that personal data is handled responsibly, and individuals' privacy rights are respected. Moreover, accountability and transparency are essential ethical considerations. Diverse professionals must be diligent in making AI systems explainable and auditable to avoid potential negative consequences. Additionally, there is a concern about the impact of AI on employment.

Professionals need to consider the potential displacement of jobs and work towards minimizing adverse effects on individuals and communities. Lastly, there is the broader ethical concern of power and control. AI technology should not concentrate power in the hands of a few or reinforce existing inequalities. By acknowledging and addressing these ethical concerns, diverse ICT professionals can work towards developing AI systems that align with societal values, promote fairness, and have a positive impact on individuals and communities.

Identifying different views, theories and perceptions in the AI ethics discussion could improve the potential of AI technologies on a global scale in a multidisciplinary social, cultural, political and ethical manner. In literature, a number of research initiatives and academic endeavours targeted to identify unacceptable risks and prohibited AI practices. The challenging point is which categories of high-risk AI systems have been elaborated so far, what redress mechanisms are revoked and the opening issues by diverse fields, sectors and environments.

Given the main motivation for AI’s use and its relevant applications, which is the economic benefits and sustainability for different sectors such as education, healthcare, business management and agriculture, it is of great importance to (a) review the perceptions of diverse actors and environments in AI world and (b) stress the achievements and the gaps so far in respect to ethical guidelines’ implementation. The present work, therefore, reviews relevant initiatives such as the [masked due to blind review] project and [masked due to blind review], that attempted to converge different contexts on this topic in order to acquire sufficient

evidence for effective mechanisms, strategies and policies. In addition to this, and in order to prepare a more consolidated work, interviews with companies including diverse ICT professionals in the use and implementation of an AI-based solution will be conducted. These interviews aim to discuss the ethical issues that might be raised during these processes, and how they are handled.

The present work highlights also the dynamics and interactions that could be deployed between diverse AI actors and stakeholders and investigates if there is balance and complete consideration of AI ethics in a horizontal way. Also, it leverages the synergies of research initiatives and the tools that these synergies use for a dynamic and interactive knowledge diffusion. AI technology can be influenced by those "who build it and the data that feeds it" (Kim, 2017). Therefore, the role of context, education and culture could be reflected in AI development and use and vice versa. Upon this, among the considerations of the present work is how sustainability in education and training programs of ICT professionals can be achieved and which pathways and what kind of effort and individual involvement are required to meet AI challenges.

This attempt is currently happening at an official level, in terms of the EU Legislation which has been voted, the Artificial Intelligence Act. The Commission proposes to establish a technology-neutral definition of AI systems in EU law and to lay down a classification for AI systems with different requirements and obligations tailored to a 'risk-based approach (Madiega, 2021).

KEYWORDS: AI technologies, ethical and security controversies, inclusive and user-friendly AI, cultural and social diversity, ICT professionals, Law Enforcement Agencies.

REFERENCES

Carvalho, L., Martinez-Maldonado, R., Tsai, Y. S., Markauskaite, L., & De Laat, M. (2022). How can we design for learning in an AI world? Computers and Education: Artificial Intelligence, 3, 100053.

Eurostat, Use of artificial intelligence in enterprises, (2022), https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., ... & Tang, J. (2022). Trustworthy ai: A computational perspective. ACM Transactions on Intelligent Systems and Technology, 14(1), 1- 59.

Meyer, D., & Henke, M. (2023). Developing design principles for the implementation of AI in PSM: An investigation with expert interviews. Journal of Purchasing and Supply Management, 100846.

Kim, P. T. (2017). Auditing algorithms for discrimination. *U. Pa. L. Rev. Online*, *166*, 189. Madiega, T. A. (2021). Artificial intelligence act. *European Parliament: European Parliamentary Research Service*.

USERS' PERCEPTION OF PRIVACY BOUNDARIES IN THE DIGITAL WORLD: A STUDY FROM THE ARAB WORLD

Ala Ali Almahameed, Mario Arias-Oliva, Jorge Pelegrín-Borondo, Mar Souto-Romero

Social and Business Research Lab, Universitat Rovira i Virgili (Spain), Management and Marketing Department, Complutense University of Madrid (Spain), Economics and Business Department, University of La Rioja (Spain), School of Business and Communication, Universidad Internacional de La Rioja (Spain)

a.mahameed82@gmail.com; mario.arias@ucm.es; jorge.pelegrin@unirioja.es; mar-souto@gmail.com

EXTENDED ABSTRACT

The development of the internet and social media has dramatically altered the way people communicate and share information, creating new opportunities for social interaction, business, and entertainment. In fact, social media platforms like Facebook, Twitter, and Instagram allow people to communicate with each other in a real-time, despite their physical location. Furthermore, social networks have expanded and diversified their offerings. For example, Facebook has acquired Instagram and WhatsApp, and it has launched features such as Facebook Live and Facebook Marketplace. Likewise, Snapchat has introduced new features such as Snap Map and augmented reality filters. As well, LinkedIn has introduced new tools for job seekers and recruiters, and Twitter has expanded its focus on news and live events. In general, social media platforms make it easy to share news, articles, photos, and videos with friends, family, and followers (Dizikes, 2020). Overall, social media networks have evolved to become an integral part of daily life for many people, with a wide range of uses and features. According to Smart Insights, the number of social media users globally increased from 4.2 billion in January 2021 to 4.62 billion in January 2022. As a result, people post 500 million tweets, share over 10 billion pieces of Facebook content, and watch over a billion hours of YouTube video during the day (Chaffey, 2023).

There has been a significant growth in using social media networks in the Arab world over the past few years. According to a report by Global Media Insight (2023), for instance, there are over 28.8 million active social media users in Saudi Arabia, which represents 79.3% of the total population, and 99.8% of the UAE population is using social media networks. Moreover, Kemp's (2023) report shows that 96.8% of Qatar's population, 58.8% of Jordanians, and 47% of Egyptians are using social media networks. This growth in social media usage can be attributed to various factors, including the development of ICT, the increasing availability of affordable smartphones, high internet penetration rates, and the growing popularity of social media platforms among Arab youth (Alammary 2022). Same as the rest of the world, Arabs are using social media networks for different purposes, such as communicating with friends and relatives, shopping, seeking jobs, and so on. For example, social media networks are widely used by elites and everyday citizens to discuss politics and achieve political goals. In this context, a study by National Endowment for Democracy found that social media has become a powerful tool for political mobilization in the Arab world. Researchers have also used social media data to study political behaviour in the Arab world (Siegel, 2019). Furthermore, a report by Pew Research

Center claimed that social media played a role in the Arab uprisings that began in 2010 (Brown, Guskin & Mitchell, 2012).

Despite that social media has become an integral part of our lives, it comes with its own set of privacy concerns. Some of the most common social media privacy issues include social media phishing scams, hacking and account takeovers, shared location data used by stalkers and predators, data mining leading to identity theft, privacy “loopholes” exposing your sensitive information, employers or recruiters evaluating you based on your posts, doxing leading to emotional distress or physical harm, cyberbullying and online harassment. Furthermore, social media platforms such as Facebook, Twitter, and Instagram collect and store massive amounts of personal data from users, including their location, search history, and social interactions (Zhang et al., 2020). This data is used to deliver personalized content and advertising to users, which can be beneficial for some individuals. However, concerns arise when this personal data is misused, shared without consent, or exploited for profit. For instance, millions of Facebook users' data was harvested without their consent and used for political advertising (Cadwalladr & Graham-Harrison, 2018). Moreover, Children are at risk of online grooming, cyberbullying, and exposure to inappropriate content, while individuals with disabilities may be more susceptible to online scams and phishing attacks (Kargupta & Kumar, 2021).

The basis for morality and ethics in the Arab world, especially for Muslims, is primarily derived from the Qur'anic text and the verbatim quotes from the Prophet Muhammad, known as the Sunnah.

These sources constitute the foundation of Sharia law, which not only shapes the judicial system but also establishes societal norms and expectations for behaviour. The concept of privacy is highly valued and is an integral part of daily life in the Arab world. The Holy Quran emphasizes the importance of seeking permission before entering someone's home as a means of safeguarding privacy and maintaining the sanctity of the house and body. The act of knocking on a door three times before entering is intended to prevent unintentional intrusion on one's private space, especially in situations where one may be in a state of undress or with their spouse or family. Failing to seek permission and entering without consent can lead to an invasion of privacy (Norah & Sarah, 2016).

The Arab world has a unique cultural and social context that affects the way people view privacy. For instance, people in the Arab world may value privacy differently than people in the Western world. Understanding these cultural differences is crucial in designing effective privacy policies that are sensitive to the needs and expectations of the Arab population (Askool, 2013). Besides, studying social media privacy concerns in the Arab world is required to understand cultural differences, political implications, business opportunities, and human rights issues. It is essential also to develop effective privacy policies and protect the privacy of individuals in the region (Norah & Sarah, 2016). Furthermore, social media has played a crucial role in the Arab Spring uprisings that took place in the region. These events have highlighted the importance of social media platforms as tools for political mobilization and expression of dissent. In fact, privacy concerns in the Arab world are not just about protecting individual rights, but they also have significant political implications (Abokhodair et al., 2017).

The Arab world is a rapidly growing market for social media platforms, with a high rate of social media adoption among its population. Understanding privacy concerns in the region is crucial for social media companies that wish to tap into this market and build trust with their users (Khawla F Ali et al., 2020). Also, privacy is a fundamental human right, and social media privacy

concerns in the Arab world are no exception. In this context, the previous research focused on the effect of cultural restrictions on individuals' motivation, users' attitudes, intentional behaviour, and social media's actual use, in addition to understanding the purposes, benefits, and risks of its use (e.g. (e.g., Askool, 2013; Abaido, 2020; Asiri et al., 20217). Also, some of the previous research investigated the role of Islam and cultural traditions in constructing norms around privacy (e.g., Abokhodair et al., 2017; Shehu et al., 2017). However, there are limited studies that investigate the impact of culture and governing laws in mitigating the negative impact of privacy while using social media websites. In particular, understanding and respecting the privacy boundaries of other users while interacting with them on these platforms. Hence, the extended research aims to investigate the role that morality and ethics that are driven from Islam and Arab culture are playing in regulating users' interaction with others over social media websites if associated with national laws that govern such interaction. In this way, the researchers believe that the research results will introduce a practical solution that could be used to make social media platforms a safe place for users, especially while interacting with others. Also, the extended research will propose recommendations for future research to expand the study and generalize its results.

KEYWORDS: Social Media, Privacy, Arab World, Ethics.

REFERENCES

A.Siegel, Alexandra (2019). Using Social Media Data to Study Arab Politics. *APSA MENA Politics*. Retrieved from https://apsamena.org

Abaido, G. M. (2020). Cyberbullying on social media platforms among university students in the United Arab Emirates. *International journal of adolescence and youth*, *25*(1), 407-420.

Abokhodair, N., Abbar, S., Vieweg, S., & Mejova, Y. (2017). Privacy and social media use in the Arabian Gulf: Saudi Arabian & Qatari traditional values in the digital world. *The Journal of Web Science*, *3*.

Abokhodair, N., & Vieweg, S. (2016, June). Privacy & social media in the context of the Arab Gulf. In *Proceedings of the 2016 ACM conference on designing interactive systems* (pp. 672-683).

Alammary, J. (2022). The impact of social media on women's empowerment in the Kingdom of Bahrain. *Gender, Technology and Development*, 26(2), 238-262.

Ali, K. F., Whitebridge, S., Jamal, M. H., Alsafy, M., & Atkin, S. L. (2020). Perceptions, knowledge, and behaviors related to COVID-19 among social media users: cross-sectional study. *Journal of medical Internet research*, *22*(9), e19913.

Asiri, E., Khalifa, M., Shabir, S. A., Hossain, M. N., Iqbal, U., & Househ, M. (2017). Sharing sensitive health information through social media in the Arab world. *International Journal for Quality in Health Care*, *29*(1), 68-74.

Askool, S. S. (2013). The use of social media in Arab countries: A case of Saudi Arabia. In *Web Information Systems and Technologies: 8th International Conference, WEBIST 2012, Porto, Portugal, April 18-21, 2012, Revised Selected Papers 8* (pp. 201-219). Springer Berlin Heidelberg.

Brown, Heather, Guskin, Emily, & Mitchell, Amy (2012). The Role of Social Media in the Arab Uprisings. *Pew Research Center*. Retrieved from https://www.pewresearch.org

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The guardian*, *17*(1), 22.

Chaffey, Dave (2023, June 07). Global Social Media Statistics Research Summary 2023. *Smart Insights.* Retrieved from *https://smartinsights.com*

Dizikes, Peter (2020, September 24). Why social media has changed the world — and how to fix it. *MIT News Office*. Retrieved from https://news.mit.edu

GMI Blogger (2023, March 06). Saudi Arabian Social Media Statistics 2023. *Global Media Insight*. Retrieved from https://globalmediainsight.com

Kemp, Simon (2023, January 28). Digital 2023 Deep-Dive. *DATAREPORTAL*. Retrieved from https://datareportal.com

Shehu, M. I., Othman, M. F. B., & Osman, N. B. (2017). The social media and Islam. *Sahel Analyst: Journal of Management Sciences*, *15*(4), 67-80.

Zhang, D., Zhou, L., & Lim, J. (2020). From networking to mitigation: the role of social media and analytics in combating the COVID-19 pandemic. *Information Systems Management*, *37*(4), 318-326.

ETHICAL INFLUENCERS: THE RIGHT PATH FOR DIGITAL INFLUENCERS

Orlando Lima Rua, Rafaela Mesquita, António Oliveira

Center for Organisational and Social Studies of the Polytechnic of Porto – CEOS.PP (Portugal), Porto School of Accounting and Business – ISCAP-P.PORTO (Portugal)

orua@iscap.ipp.pt; 2201984@iscap.ipp.pt; ajmo@iscap.ipp.pt

EXTENDED ABSTRACT

With the growth of the internet, consumers are using social media more for information and enlightenment on which to base their buying decisions about a particular product or service. Social networks, integrated into the daily habits of many of their users, provide the consumer access to information about both products and brands, influencing the consumer's decision process and, consequently, changing the way brands communicate with consumers (Castelo, 2020, pp. 32-33). For Castelo et al. (2020), social networking applications are applications that allow users to connect with each other by creating profiles of personal information and exchanging messages with each other, and to share photos, videos, and audio files. Also according to these scholars, (1) the possibility of connecting with other users and sharing information and opinions about products and brands has made these applications an effective vehicle for word of mouth, and (2) by allowing a constant exchange of information between users, social networks are seen as both a challenge and an opportunity for brands, and marketers are forced to explore the feasibility and possibility of integrating these applications into their strategies to communicate more effectively with their consumers and strengthen relationships with them.

It is predicted that mobile usage will grow globally with direct impact on content creation by digital influencers. In the influencer marketing industry, one can see the emergence of entrepreneurs who have turned their influence into a business by launching their own products. digital influencers in looking to partner with brands to invest in a long-term relationship as it not only builds trust but adds legitimacy, another trend that is growing over time is authenticity in media. Social media communication is no longer robust because of the great diversity of digital platforms, giving users the power to be active and present in the communication process with brands. With this, the main purpose of social networks is specifically to empower people to publish content on the internet, such as photos, posts, videos, among others, all over the world (Marques, 2022).

Influencers are referred to as opinion leaders, meaning that they are personalities with a large network of followers and fans, who have become dominant members of this online community (Casaló et al., 2020). In this perspective, digital influencers have been gaining the attention of brands due to (1) the increasing dissemination of their content (Kim & Park, 2023) and (2) the mitigation of risks associated with human error in marketing campaigns, as they are ageless, digital avatars with no offline existence that could potentially jeopardise their online persona (Kim & Wang, 2023).

It is therefore important to delimit the ethical issues associated with digital influencers, taking into account the potential for deception and issues linked to moral responsibility (Kim & Wang, 2023, p. 4), such as "Transparency about the identity of these influencers and their content

sources is paramount, underscoring the need for disclosure about those responsible for their creation and management.". Besides, it is also relevant to research how can digital influencers leverage, by the use of their social media networks, consumption-driven social change linked to ethical consumption (Aboelenien et al., 2023). Therefore, the concept of ethical influencer arose. Also according to these scholars, "legitimate their accounts via close-up of personal practices, as opposed to an articulated persona, and connect with divergent audiences to advocate for the need change." (p. 1).

For Lou and Yuan (2019), digital influencers are online personalities who have gained a large number of followers, through one or more social networks (e.g., Facebook, Instagram, Twitter, Tik Tok, YouTube, or personal blogs), who have a great influence on their followers and unlike public figures, influencers are "regular people" who create content in specific areas such as healthy living, travel, food, lifestyle, beauty, or fashion. The strategy that influencers use to capture attention for a particular target is to convince them that what they believe, or think is based on false information or otherwise may convince them that what "other people" believe or think they know is based on false information, leading to a feeling of superiority. Which leads influencers to learn how to capitalize on the opportunities the Internet offers to shape a reality that is available to its users (Forest, 2021).

The consumer may turn to one source about fashion, while being inclined to look for another about cooking, and so on. For example, by following a nutritionist on Instagram, the consumer will be influenced, but only in the context of healthy eating and other topics related to active living, tips and eating. This means that certain figures are influential in their area of expertise. For Levin (2020), there are three levels of influence: (1) experience and credibility are at the first level, (2) on the second level is the strength of the relationship with the followers and the trust they have in the influencer they follow; the better an influencer knows his followers and the greater the trust, the more targeted and effective his content is and (3) the number of followers you can reach is the third level of influence; by reconciling and optimizing these three levels, an influencer will be able to produce quality content and effectively advertise any product or service.

Typically, influencers have the power to try new products or services according to their domain of interest earlier than most consumers, which gives them early insight into how these products and services fit into their lifestyles. Influencers leverage this early insight to review products, make recommendations, and offer tips to their followers, thus allowing them to build credibility and monetize their work through partnerships and campaigns they run. Followers perceive influencers to be popular personalities and more trustworthy than celebrities, as they create a connection with their followers (Wondwesen & Wood, 2021).

For Lou and Yuan (2019), social media influencers are online personalities with many followers, across one or more social media platforms, who have an influence on their followers. Contrary to celebrities who are well-known via traditional media, influencers are "regular people" who have become "online celebrities" by creating content on social media.

Influencers perform marketing activities through advertising. Influencers promote brands or products through their content. In short, influencers can successfully perform marketing activities by introducing the product as organic content versus commercial content. Social media users are more likely to be receptive to a promotional message when it is perceived as a genuine message from the influencer. With this, they concluded that, content that matches an

influencer's domain of interest, generates a more favorable evaluation of the products they sponsor (Kim, 2020).

Forest (2021) uses the term "digital influence warfare" to refer to a form of psychological persuasion whereby the influencer can manipulate the believes and behaviors of others. This can be with the use of persuasive tactics, like information and disinformation, provocation, identity deception, computer network hacking, altered videos and images.

Today, society is addicted to technology and social media, which is the dream of the younger generation that they want to be famous digital influencers. Digital influencers want to engage with customers in a more personal, mobile, and social way. As a result of that, brands trust more on digital influencers for continued sales growth and conversion rates (Teixeira, 2022). According to, Wang and Huang (2020) about 80% of marketers believe that digital influencers are powerful facilitators of consumer engagement and purchase. Digital influencers are individuals whose personal social media accounts have a stable and high number of followers. For example: a nano-influencer on Instagram has 1,000 to 5,000 followers, a micro-influencer has 5,000 to 20,000, a mid-level influencer has 20,000 to 100,000, a macro-influencer has 100,000 to a million, a mega-influencer has more than a million. Finally, 60% of brands focus on influencer strategies when increasing investments in social commerce.

With the growing relevance of digital influencers in peer-to-peer relationships and the potential adverse effects associated with idealised body representations, there are ethical implications about their use in influencer marketing that need to be investigated in the future (Kim & Wang, 2023). As brands and businesses optimise social spaces and networks, consumers must look beyond the interests and commitments of influencers, because ethical influencers do not accept sponsorship in order to preserve their legitimacy (Aboelenien et al., 2023; Schouten et al., 2020). The right path for digital influencers to satisfy consumer needs is by transforming them into ethical influencers.

KEYWORDS: Ethical influencers, digital influencers, social media, ethical consumption, moral responsibility.

REFERENCES

Aboelenien, A., Baudet, A., & Chow, A.M. (2023). 'You need to change how you consume': etichal influencers, their audiences and their linking strategies. *Journal of Marketing Management*. https://doi.org.10.1080/0267257X.2023.2218853

Casaló, L.V., Flávia, C., &. Ibañez-Sánchez, S. (2020). Influencers on Instagram: Antecedents and consequences of opinion leadership. *Journal of Business Research*, *117*, 510-519. https://doi.org/10.1016/j.jbusres.2018.07.005

Forest, J. (2021). *Digital Influence Warfare in the Age of Social Media.* California: ABC-CLIO.

Kim, K. &. (2020). Influencer advertising on social media: The multiple inference model on influencer-product congruence and sponsorship disclosure. *Journal of Business Research*, *130*, 405-415.

Kim, H., & Park, M. (2023). Virtual influencers' attractineness effect on purchase intention: a moderated mediation model of the Product-Endorser fit with the brand. *Computers in Human Behavior*, *143*, 107703. https://doi.org/10.1016/j.chb.2023.107703

Kim, D., & Wang, Z. (2023). The ethics of virtuality: navigating the complexities of human-like virtual influencers in the social media marketing realm. *Frontiers in Communication*, *8*, 1205610. https://doi.org/10.3389/fcomm.2023.1205610

Levin, A. (2020). *Influencer Marketing for Brands: What YouTube and Instagram Can Teach You About the Future of Digital Advertising.* Sweden: Apress.

Lou, C., & Yuan, S. (2019). Influencer Marketing: How Message Value and Credibility Affect Consumer Trust of Branded Content on Social Media. *Journal of Interactive Advertising*, *19*(1), 58-73. https://doi.org/10.1080/15252019.2018.1533501

Marques, V. (2022). *Marketing Digital 360.* Digital 360.

Schouten, A.P.; & Janssen, L., & Verspaget, M. (2020). Celebrity vs. Influencer endorsements in advertising: The role of identification, credibility, and product-endorser fit. *Frontiers International Journal of Advertising*, *39*(2), 258-281. https://doi.org/10.1080/02650487.2019.1634898

Teixeira, P.M. (2022). *Digital marketing trends.* Porto: CEOS.

Wang, P., & Huang, Q. (2020). Digital influencers, social power and consumer engagement in social commerce. *Internet Research*, *33*(1), 178-207. https://doi.org/10.1108/INTR-08-2020-0467

Wondwesen, T, & Wood, B.P. (2021). Followers' engagement with instagram influencers: The role of influencers' content and engagement strategy. *Journal of Retailing and Consumer Services*, *58*, 102303. https://doi.org/10.1016/j.jretconser.2020.102303

HUMAN-CENTRED ARTIFICIAL INTELLIGENCE, DISRUPTION, AND EXPLAINABILITY

Philip J. Nickel

Eindhoven University of Technology (Netherlands)

p.j.nickel@tue.nl

EXTENDED ABSTRACT

Artificial intelligence (AI) is designed to take over human tasks. When tasks that originally required human capacities are completely automated by AI, AI fully displaces humans in the exercise of the relevant capacities to complete that task. Let us call this the AI-displacement pattern. This can be disruptive because it forces humans to find new tasks or competencies, satisfying their basic needs in new ways. Some models of automation focusing on levels of control assume that this is the central pattern (Parasuraman et al., 2000). This feeds a view on which the main ethical concerns around AI-based automation are job displacement and loss of meaningful work. Such concerns have been addressed by proposals for retraining, by assurances that AI creates as many jobs as it replaces, and by discussions around the meaningfulness of the new tasks that replace old ones (Smids, Nyholm, & Berkers 2020). Discussing a specific case of AI-based analysis of protein folding, Bankins & Formosa argue that "While AlphaFold can assume significant tasks previously done by human scientists (i.e., determining protein structures) this should positively impact, or at least have a neutral effect, on task integrity if it allows scientists to re-focus their work efforts on other important aspects of their broader goal of curing diseases. However, there remain risks to AI being used in this way. Continuing with this example, if scientists have trained for many years to do the experimental work that AlphaFold can now do more quickly and accurately, this generates significant risks for their ability to exercise their full capacities, demonstrate their mastery, and utilise the skills they have invested years in developing to reach their full potential" (2023).

Yet many upcoming AI applications are "human-centred" AI (HCAI) and therefore do not neatly fit the AI-displacement pattern. They involve human-AI interaction or human-in-the-loop solutions, where humans are expected to interact with, evaluate and conditionally override AI-supported actions or judgments in the execution of tasks (Schneiderman 2020). The AI does not fully displace humans in the exercise of the task, and it does not fully displace the exercise of the relevant human capacities for executing that task. Instead, there are high levels of automation *and* high levels of human autonomy and control in relation to a task that is more or less similar to the original task. Usually this is argued to be a positive development in the application of AI (ibid.). To some extent, retraining and care for the preservation of meaningful work can address concerns about the HCAI pattern as well. However, there is a need for a closer look at this alternative pattern to discern any distinctive ethical concerns that it may raise. It has been argued that the application of HCAI is a "third path" that can result in amplifications of human autonomy and other meaningful aspects of work (Bankins & Formosa 2023). In this third path, "AI is neither assuming specific tasks that a human previously did (as in the first path) nor does managing the AI constitute a worker's primary role (as in the second path), but rather the technology assists the worker to do her existing work better" (ibid.).

This paper aims to complement the account of Bankins & Formosa by analysing the relevant HCAI pattern and their "third path" in terms of disruption of human discretion, where this latter concept is understood as a kind of authority to make a judgment within a domain. Although existing work may be done differently, interaction with AI raises questions about whether the task is done "better" from the human user's point of view. This analysis builds on theories of socially disruptive technologies (Baker 2013, Hopster et al. 2022) and of discretion (Dworkin 1977). Even when it works well, HCAI is likely to result in well-grounded perplexity about whether and when to trust or rely on AI in the execution of a task, due to legitimate user concerns about responsibility gaps, hidden political and institutional agendas, and technological dependency. Such an account is different from Bankins & Formosa in that it is not psychological but moral-epistemological: it relates to legitimate moral concerns that a user may have about relying on HCAI when first experiencing it. It raises doubts about Bankins & Formosa's prediction that this path generally leaves intact the elements of meaningful work — integrity, skill cultivation, and task significance — or even improves them. HCAI creates legitimate perplexity, leading to disruption in the exercise of human discretionary authority. Accounts of disruption as moral uncertainty point to the hazard and potential harm associated with it, as well as the challenge that it poses for moral agents (Author Reference).

Explanatory AI (XAI), considered as part of the HCAI paradigm (Schneiderman 2020), can help to bolster trust and mitigate feelings of disruption (Bankins & Formosa 2023). However, it is argued here that in doing so it might simply beg the question of trust or distract humans from the underlying concerns. This is because even when functioning well, XAI does not actually answer question about what counts as responsible agency on the part of the user and what kinds of dependency are problematic, nor does it reveal all the hidden political and institutional agendas that may be involved when introducing an AI application and asking users to rely on it. The upshot is that HCAI may be *morally* disruptive in a different way than non-human-centred AI that fully takes over tasks. These moral disruptions might also be experienced by a much wider range of people than before, because of the dilemmas raised by widespread availability and experimentation with content generation technologies such as natural language generation using large language models (e.g., Chat-GPT), whose application can fit neatly within the "third path" and the HCAI paradigm.

After a review of the literature on HCAI and the arguments for adopting it in Section 2, the paper applies the idea of moral disruption to it in Section 3. It is then argued in Section 4 that explanatory AI cannot fully mitigate this disruption. The paper closes with some observations about the future of HCAI.

KEYWORDS: Human-centred artificial intelligence, discretion, moral disruption, explainable AI, meaningful work.

REFERENCES

Baker, R. (2013). *Before Bioethics*. Oxford University Press.

Bankins, S. & Formosa, P. (2023). The Ethical Implications of Artificial Intelligence (AI) for Meaningful Work. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-023-05339-7

Dworkin, R. (1977). *Taking Rights Seriously*. Harvard University Press.

Hopster, J.K.G., C. Arora, C. Blunden, C. Eriksen, L.E. Frank, J.S. Hermann, M.B.O.T. Klenk, E.R.H. O'Neill, & S. Steinert. (2022). Pistols, Pills, Pork and Ploughs: The Structure of Technomoral Revolutions, *Inquiry,* http://doi.org/10.1080/0020174X.2022.2090434.

Parasuraman, R., Sheridan, T.B., Wickens, C.D. (2000). A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man and Cybernetics---Part A: Systems and Humans*, 30:3, 286-297.

Schneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, 36 (6), 495-504, http://doi.org/10.1080/10447318.2020.1741118

Smids, J., Nyholm, S., & Berkers, H. (2020). Robots in the Workplace: a Threat to – or Opportunity for – Meaningful Work? *Philosophy & Technology* 33: 503-522.

4. Reducing Gender Gap as We build an Inclusive Community within a Smart City

Shalini Kesar, Southern Utah University; Graciela Padilla Castillo, Complutense University of Madrid

DIVERSITY MISSING IN CYBERSECURITY

Shalini Kesar

Southern Utah Univeristy (USA)

kesar@suu.edu

EXTENDED ABSTRACT

This paper motivation is from comment in a comment in the Women in Cybersecurity (2022); "We know that the representation of women in cybersecurity hovers around 24%, far lower than it should be," commented Lynn Dohm, executive director of WiCyS, in a statement. "We wanted to find out why this was the case and were somewhat — but not entirely — surprised that the most common source of women's feelings of exclusion came from people, not company policies. This highlights the fact that we still have a long way to go when it comes to accepting women in the cybersecurity industry." This paper discusses and possible solution to reduce the diversity gap in the cybersecurity field. It presents the reality check of cybersecurity that clearly highlights lack of women and underrepresented groups in this field. It also reflects on the various efforts proposed or/and implemented to address this concern. This is ongoing research where the author is passionate and motivated to address as well as propose some pragmatic steps towards creating a cybersecurity pipeline with diversity including women and underrepresented groups.

Many research and article throw light on the underlying factors of lack of diversity in STEM field including cybersecurity and the impact of stereotypes and gender bias. Hundreds of studies have conducted research on, for example, the power of stereotypes to influence performance through a phenomenon known as "stereotype threat.", disengagement from fields in which women are negatively stereotyped, such as computing and cybersecurity (solving the equation). The good news is that practitioners, academia and community as whole acknowledges that there are concerns in the talent pool or the lack of talent pool that goes beyond just demand and supply gap numbers in the cybersecurity area. The Economics Forum stated two major issues: 1) The global cybersecurity skills gap; and 2) The lack of diversity in the cybersecurity workforce. Studies in SANS (ICS, 2023) highlight that 3.4 million people are needed to fill the global cybersecurity workforce gap. Consequently, a survey by the World Economic Forum (2022)) that 59% of businesses would find it difficult to respond to a cybersecurity incident due to the shortage of skills. It also states that the cybersecurity sector needs 3.4 million people to fill its workforce gap.

There are many articles that reflect on how to create a pipeline starting from kindergarten school. This can be empowering and can perhaps impact the shift in the mindset of a field itself. For example, the author the author's white paper, collaborative paper with Microsoft (2018) highlights how engaging young girls creating a role and having hands-on activities can spark interest in cybersecurity and STEM fields.

This paper focuses on some of the pragmatic ways the organization 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. This paper takes the four categories from her findings of the research at K12 into the context of a workplace environment. It uses the "9 Strategies to

Improve Gender Diversity in the Security Workforce" Security Intelligence, 2020) article as a starting point to highlight some examples to retain women in this ever evolving field. 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 Advancement. Support Competitions and Scholarships Specifically for Women. These are explained below.

Support Competitions and Scholarships Specifically for Women: This refers to various scholarships or events that are inclusive to young women. For example. host a security-focused hack-a-thon or a capture the flag competition specifically for women that focuses on hands-on security skills, teamwork and applications to real-world cybersecurity challenges. It is also a good idea to share opportunities about scholarships and competition that are women centric conferences. For example, Women in Cybersecurity annual conference or The Women's Society of Cyberjutsu (WSC), a 501(c)3 non-profit, is dedicated to raising awareness of cybersecurity career opportunities and advancement for women in the field, closing the gender gap and the overall workforce gap in information security roles. The use of inclusive language in hiring efforts and involving women in recruitment process are important factors to encourage women to consider cybersecurity career. For example, advertisements in cybersecurity positions with language and images that are inclusive of all applicants can motivate women to apply for employment. Another example of this strategy is to involve senior-level women directly in the interviewing and recruiting processes. This makes the applicants aware there are other women in the organization as well as opportunities for them in advancement within the organization.

Other strategies are to provide opportunities for lateral growth and enable employees to pursue external certifications. Creating professional development programs for new hires in cybersecurity that allow them to rotate through different areas of the organization that deal with security. This can help them determine which areas they are most interested in and where they might find the best fit in the long term. In this field, certifications add value to the professional development. Hence, providing support for women to engage in external training and certification programs related to cybersecurity, such as Certified Information Systems Security Professional (CISSP) training or Certified Information Security Manager (CISM) certification. The other two strategies in this paper refer to considering women who are rejoining the workforce and also offer fair and equitable Compensation. It is critical that salaries across cybersecurity roles ensure that women are not being paid less than men for the same job. Finally, it is important to organize pathways for advancement. The author has experienced that first hand as she has conducted many outreach projects for high and middle school girls that are linked to STEM including cybersecurity subjects. In the article "Empowering women can help fix the cybersecurity staff shortage" (2022) published in The Economic Forum states that Our survey corroborated some traditional thinking – but refuted other key, long-held hypotheses: 1) It's important to engage girls in STEM early. Their research confirmed this hypothesis as a majority – 78% – of our respondents said that they had first developed an interest in STEM in middle school or high school.

Women are aware of cybersecurity. There's a perception that awareness of cybersecurity is low among women. We found the opposite to be true: 82% of survey respondents said they had some or a lot of knowledge of cybersecurity; 2) Women have access to cybersecurity education. Another perception: low participation of women in cybersecurity because they lack access to

cybersecurity education. Our survey indicated otherwise. Specifically, 58% of respondents said they had access to cybersecurity education, and 68% had already taken a cybersecurity-related course; 3) Role models and senior encouragement are critical. That's what anecdotal evidence suggested, and our survey validated the hypothesis. Role models played an important factor to avoid the negative perceptions of cybersecurity as a career choice. The top three priorities for women in choosing a job are contributing to society, earning a high salary and having a good work-life balance. However, 37% of respondents regard cybersecurity as a field where achieving that balance is difficult; 4) Lack of awareness also had a negative perceptions with a mindset that in cybersecurity is that it's often regarded as a "boys' club".

In light of the above, many articles highlight the importance of reducing the gap in cybersecurity field. For example, in the Cybercrime Magazine, Osborne (2022) 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 (Panhans et al, 2022) 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".

KEYWORDS: Cybersecurity, women in cybersecurity, diversity.

REFERENCES

ICS (2022). Five Startling Findings In 2023's ICS Cybersecurity Data. Retrieved from https://www.sans.org/blog/five-startling-findings-2023-ics-cybersecurity-data/

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/RE1UMWz

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/

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-a-win-win

Women in Cybsersecurity (2022), retrieved from http://www.wicys.org

World Economic Forum (2022). Empowering women can help fix the cybersecurity staff shortage. Retrieved from https://www.weforum.org/agenda/2022/09/cybersecurity-women-stem/

CYBERSECURITY AS A GOOD LIFE PATH FOR EVERYONE

Aleksandra Pyrkosz, Sabina Szymoniak

Department of Computer Science, Czestochowa University of Technology (Poland)

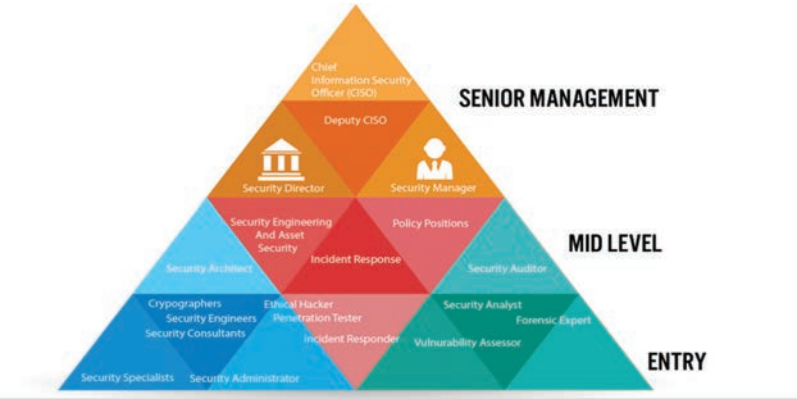
aleksandra.pyrkosz98@gmail.com; sabina.szymoniak@icis.pcz.pl

EXTENDED ABSTRACT

Cybersecurity is one of the most exciting areas of work and science. It is evident in the digital area, where almost every company, regardless of size, has an Internet connection. Access to the network offers many opportunities but also entails new challenges. One of them is the proper care of security in cyberspace.

Along with the growing importance and use of digital technologies, the number of threats that organisations must counter also increases. Cybercriminals are using more advanced attack methods, so building security is evolving and improving. The introduction of effective security practices in cyberspace is crucial for protecting information, maintaining customer trust, and maintaining the stability of the company's operation in the era of universal digitisation. Responsible for cybersecurity is a continuous process and requires constant monitoring, adapting strategies and protective measures to the changing threat landscape. This is a complicated process, but there are different ones on the market with the possibility of implementing certain facilitations while maintaining an appropriate level of safety ((Steingartner et al., 2022), (Nwankpa & Datta, 2023)). Figure 1 summarizes the cybersecurity career path and shows how different activities cybersecurity specialities perform.

Figure 1. Cybersecurity career path.



- Source: https://www.spiceworks.com/tech/it-careers-skills/articles/cybersecurity-career-path/

Moreover, cybersecurity issues make a perfect space for researchers. The computer systems exposed to cyberattacks need specially designed algorithms and techniques for security improvement. Such systems need secure communication between network nodes. Thus, researchers must propose new security protocols that will define the sequence of the steps

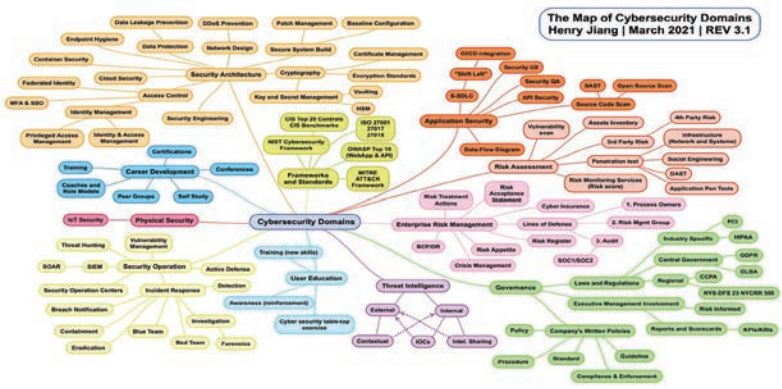
during the communications and use security techniques like encryption, timestamps, pseudonymity or hashing functions. Security protocols can be designed for cross-domain or specific solutions. Also, the security protocols should be constantly verified to check if they provide an appropriate security level ((Bartłomiejczyk et al., 2022), (Szymoniak, 2021)).

Next, computer systems need tools to check the system's vulnerabilities and Intrusion Detection and Prevention Systems (IDPS). IDPSs monitor computer networks to detect and respond to suspicious or harmful activities. They analyse network traffic and identify patterns that may indicate unauthorised access attempts, attacks, or other unwanted activity. IDPSs monitor the network traffic, analyse the packet signatures, detect anomalies and respond to the incident. These tools mainly use artificial intelligence methods and techniques in their work (Apruzzese et al., 2023). Note that IDPSs have their limitations. They may not detect new or advanced attacks whose signatures are unknown. Therefore, IDPS systems must be updated frequently and have access to up-to-date signature databases. In addition, IDPSs can generate false positives, especially in complex networks, requiring further verification and analysis by network administrators.

Moreover, the attackers are not idle. They are constantly exploring computer systems to find new doors to get in. They improve their knowledge of systems and their hacking skills. We can indicate many cyberattack types like spoofing, known session-specific temporary information or replay attacks. Also, the performed attacks can be a combination of some typical attacks. Cyber attacks can have various effects that depend on the type of attack, its purpose and how it is carried out (data stealing, sabotage, privacy violation) (Szymoniak & Kesar, 2023).

Such a variety of possible ways to improve computer system protection gives many opportunities to choose one's life path. Figure 2 shows the map of cybersecurity domains. This diagram summarises how many opportunities and interests give us in cybersecurity. This an extensive area where everyone will find their place for work and personal development.

Figure 2. Cybersecurity domains.



Source: https://www.linkedin.com/pulse/cybersecurity-domain-map-ver-30-henry-jiang

This paper's authors also choose cybersecurity as a life path. The first met with security issues during first-degree studies. The security and network topic fascinated the author to continue the Cybersecurity speciality studies. In the MA thesis, the author considered the security

simplification process in large organizations while maintaining an appropriate security level. Also, the second author works as a Cybersecurity specialist.

The second author met with security issues during PhD studies. The author considered the verification of security protocols and the impact of time on these protocols' execution. The author obtained many exciting results and presented them in many scientific articles. Also, the second author suggests new security protocols for the Internet of Things solutions in scientific work. In didactic work, this author also focuses on security issues in a broader range. The classes concern the security of computer systems in many aspects of this topic.

Both authors met as a student and a teacher, also as a graduate student and a thesis promoter, and next as co-organizers of cybersecurity events. They would like to share their experiences connected with cybersecurity and encourage everyone to choose a similar path in life.

In this paper, the authors will share their experiences connected with cybersecurity that they received during their studies and work. They will explain why they chose such a path in life and what is so interesting and exciting in cybersecurity issues. They will show their most significant achievements. Also, they will assume challenges they faced during previous activities and perspectives for development and acquiring exciting experiences offered by cybersecurity in various directions. We believe that our experiences, insights, and tips will clarify all doubts about those unsure of choosing cybersecurity.

KEYWORDS: Cybersecurity, way of life, scientific work, experiences in cybersecurity.

REFERENCES

Apruzzese, G., Laskov, P., Montes de Oca, E., Mallouli, W., Brdalo Rapa, L., Grammatopoulos, A. V., & Di Franco, F. (2023). The role of machine learning in cybersecurity. *Digital Threats: Research and Practice*, *4*(1), 1-38.

Bartłomiejczyk, M., El Fray, I., Kurkowski, M., Szymoniak, S., & Siedlecka-Lamch, O. (2022). User Authentication Protocol Based on the Location Factor for a Mobile Environment. *IEEE Access*, *10*, 16439-16455.

Nwankpa, J. K., & Datta, P. M. (2023). Remote vigilance: The roles of cyber awareness and cybersecurity policies among remote workers. *Computers & Security*, *130*, 103266.

Steingartner, W., Možnik, D., & Galinec, D. (2022, November). Disinformation Campaigns and Resilience in Hybrid Threats Conceptual Model. In *2022 IEEE 16th International Scientific Conference on Informatics (Informatics)* (pp. 287-292). IEEE.

Szymoniak, S. (2021). Amelia—a new security protocol for protection against false links. *Computer Communications*, *179*, 73-81.

Szymoniak, S., & Kesar, S. (2023). Key Agreement and Authentication Protocols in the Internet of Things: A Survey. Applied Sciences, 13(1), 404. https://doi.org/10.3390/app13010404

INFORMING WOMEN: OVERCOMING ONLINE CHALLENGES IN POLITICAL CAMPAIGNS

Jonattan Rodríguez, Graciela Padilla-Castillo

Universidad Complutense de Madrid (Spain)

jonrodri@ucm.es; gracielp@ucm.es

EXTENDED ABSTRACT

Women's political participation faces numerous challenges in the virtual environment, where information traceability, political campaigns and misinformation play a key role. In an increasingly digitized world, access to information and women's political empowerment become vitally important. Digital platforms offer new opportunities for women to actively engage in the political sphere, express their opinions and participate in decision-making processes. However, they also pose significant challenges that should be approached with caution.

In the current landscape, digital contexts have proven to be both a source of empowerment and a fertile ground for the emergence of anti-feminist challenges and violence. Contemporary feminism has found in social networks a vital platform for mobilization and dissemination of its messages, becoming a far-reaching activist agora. However, at the same time, these digital platforms have also given rise to renewed forms of violence and opposition to the feminist movement. Delgado and Sánchez (2023) highlight in their research how the intersection between digital tools and feminism has generated both advances and challenges in the struggle for gender equality.

Added to these challenges, the growing phenomenon of disinformation and manipulation of information during electoral processes pose a threat to women's political participation. The concept of disinformation encompasses both fraudulent information content (fake news) and misleading content (misinformation), hate speech (misinformation), deliberately false speech (false speech) and unintentional misinformation by the media or journalists (missinformation). In short, disinformation involves the distortion of information through the dissemination of false news that misleads the final recipient (Rodriguez Perez, 2019, pp. 68). This spread of disinformation can undermine trust in electoral processes, influence voters' perceptions and decisions, and hinder women's active participation in politics. It is critical to address this challenge to ensure an informed and transparent environment during election periods, thereby fostering equal opportunities for women to fully participate in political processes and exercise their right to vote in an informed and informed manner.

The growing rise of fake news has revealed an increase in academic attention to the term. Faced with the landscape of so-called fake news, Rodríguez Pérez (2019) defends the use of the term disinformation, as this can encompass the multiple facets in which hoaxes, misleading or malicious content, which encompass hate speech, are propagated.

Online political campaigns are often inundated with fake news, hate speech and discriminatory narratives that can undermine public confidence in the democratic process and hinder women's active participation. It is therefore critical to develop effective strategies to combat

misinformation and promote media and digital literacy among women, giving them the tools they need to discern the veracity of information and engage in informed debates.

Likewise, online gender-based violence represents a serious obstacle to women's political participation. Online attacks, such as harassment, intimidation and defamation, can have a devastating impact on women's confidence and security, deterring them from actively participating in politics. As a result, social networks are the ideal breeding ground for those who want to attack the collective or women in a disintermediated manner. The democratization of communications that they have generated since their emergence and momentum more than 20 years ago, have allowed that there are no limits or boundaries when commenting, participating or interacting even with people we do not know.

Interconnected women are exposed to information, comments, analysis and opinions that appeal to emotionality and personal beliefs, beyond the news, giving way to the term post-truth. Those behind these publications seek to magnify, manipulate or recreate them from unreal sources. The purpose is mass dissemination and amplification through retweets, likes, or chains that go endlessly from one device to another. It is essential to adopt legislative and policy measures that address and sanction online violence, while promoting safe and inclusive environments that encourage the equal participation of all voices.

This study aims to explore the challenges and opportunities related to women's online political participation, focusing on the issues of information traceability, political campaigning, and misinformation. The research seeks to understand the impact of these factors on the creation of a strong community where women with shared technological interests can exchange ideas, identify role models, find mentors and mentees, engage in global discussions, and celebrate the power of face-to-face interactions.

To this end, the methodology employed in this study consists of a comprehensive review of case studies and analysis of relevant reports in the field. Different cases of women's online political participation will be analyzed and shared, identifying barriers, challenges and successful strategies used to overcome them. In addition, a qualitative analysis of online discourses and debates will be conducted to understand the influence of misinformation on the political process and how it affects women's participation.

The expected results of this article will provide a deeper insight into the barriers women face in their online political participation, as well as the effective strategies used to overcome these challenges. The study is expected to shed light on the importance of creating an inclusive and supportive community that promotes women's active political participation in the context of smart cities.

Discussions are expected to emerge on the effectiveness of strategies used to overcome barriers to online political participation, as well as on the responsibility of digital platforms and political actors in spreading misinformation and encouraging women's active participation. In addition, the discussion can focus on the importance of digital literacy and equitable access to technology as key enablers of women's political participation in online environments. These discussions can open up new perspectives and areas of research in the search for solutions that promote gender equality in the political sphere and address the specific challenges faced by women in the digital realm.

KEYWORDS: Women's political participation, Disinformation, Information manipulation; Electoral periods, Equal opportunities.

REFERENCES

Delgado Ontivero, L., & Sánchez-Sicilia, A. (2023). Subversión antifeminista: análisis audiovisual de la Manosfera en redes sociales. *Revista Prisma Social*, (40), 181–212. Recuperado a partir de https://revistaprismasocial.es/article/view/4958

Rodríguez Hernández, J., & Ortega Fernández, E. (2023). Contranarrativas frente a la cultura de la manósfera. @nolesdescasito y otros movimientos para frenar el odio en redes. Tyrant Lo Blanch, 197- 215.

Rodríguez Pérez, C. (2019) No diga fake news, di desinformación: una revisión sobre el fenómeno de las noticias falsas y sus implicaciones. *Comunicación*, 40, 65-74. http://doi.org/10.18566/comunica.n40.a05

ETHICS OF FEMINIST RESISTANCE AND POSSIBILITIES OF UTOPIA IN FILM: AN ANALYSIS OF THE TV SERIES *EXTRAPOLATIONS* (2023)

Asunción Bernárdez-Rodal, Ignacio Moreno-Segarra, Graciela Padilla-Castillo

Complutense University of Madrid (Spain)

asbernar@ucm.es; igmore01@ucm.es; gracielp@ucm.es

EXTENDED ABSTRACT

The ecosocial crisis we are experiencing has begun to enter the mainstream cinematic universe of adult fiction, beyond the traditional genres linked to traditional Science Fiction (Bernárdez, 2021). In recent years, stories have begun to emerge that move away from the dystopian genre, and that feed the necessary fantasy of creating a new world based on ethical values that allow us to save not only human life, but also the entire balance necessary for the life of all animals and plants on the planet (Bruna Pérez, 2020).

This research is a commitment to film creations that dare to fictionalise new forms of life and human interaction, in which nature and animal life have a place. The difficulty of film production to talk realistically about the climate crisis is a symptom of the fact that, in the capitalist system, we are not allowed to imagine a world that is not based on savage competitiveness. Fredric Jameson (2005) once said that "it is easier to imagine the end of the world than the end of capitalism". Changing the world requires changing our mental frameworks, and film and fiction are the main tools our capitalist system has to reproduce itself. In the face of what we consider leisure, amusement or entertainment, we let our guard down (Montoro Araque, 2023; Dederichs, 2023).

The media and social networks have created a particular semiosphere in which neoliberalism and competitiveness permeate everything: from formal education to popular culture, from the world of work to leisure and recreational practices. This is not just a matter of political discourses undertaken by certain conservative figures, but rather of a diffuse mentality that makes it impossible to imagine a world that functions on the basis of cooperation and radical equality. It is this mechanism that allows hundreds of climate change documentaries to be made, while new fiction remains anchored in individualistic and heroic stories (Morto, 2016; Demos, 2017; Hameed, Gunkel, & O'Sullivan, 2022).

Secondly, our work is also a bid for anti-heroism. Mainstream cinema is full of great characters (almost always male) who save other people in dramatic situations caused by climate disaster, thus reducing collective issues to individual ones. Almost all the plots of fictional productions start from the moment when a natural disaster occurs and specific people struggle hard against the risk of death. They are adventure stories, many of them framed within the genre of science fiction with all its variants. In *Waterworld* (1995) the world has flooded and the land has disappeared; in the *Mad Max* saga (1979, 1981, 1985, 2015) the main resources such as water and petrol are scarce, and this triggers a merciless fight between human beings with no empathy for each other; in *Snowpierce* (2013) the planet has frozen over and the only remaining survivors live on an eternally moving train; in *The Day After* (2004) there is a great storm that may end

civilisation; in *Interestellar* (2014) the earth's crops are destroyed, and a new planet must be found to inhabit.

Cinema is very good at imagining forms of collective death: diseases, natural catastrophes, rising sea levels, nuclear accidents, alien invasions, meteorites, genetic alterations... all of which can be seen in the film, genetic alterations... a whole panoply of disasters from which we manage to emerge triumphant thanks to sacrifice, audacity and, of course, competitiveness. Almost all these films are an exaltation of the fiercest individualism, even though, in almost all of them, there are nods to human transcendence as a whole.

All of them have at least one thing in common: they are heroic stories of people who wage a valiant struggle for survival once disaster has struck. This shifts the core of the problem: we are already experiencing the effects of global warming, and all that the film industry imagines on a massive scale are stories of how a few of us can survive? The suspicion is that, in creative circles, it is not considered "cinematic" to suggest that here and now we can do things to avoid rushing into disaster.

It seems easier to invent fantastic solutions in which the god of technology saves us as, for example, in the film *Geostrom* (2017), in which world governments unite to build a network of satellites that can control the climate, or in *A Large Life* (2017), the solution to pollution from human action on earth is to reduce the size of people to 12.7 centimetres through medical techniques. Why is there not a less spectacular approach? Why does fiction reproduce and reinforce the spectacularisation of the diseased egos of international politics and economics?

The impossibility of mainstream culture to fictionalise another possible world away from dystopias and fantastic technological solutions is the proof that we need to change our conception of the place of human beings on planet Earth. Thinking and feeling ourselves as animals among animals, as living beings among other living beings, is the only key to a peaceful future.

In our work we will analyse the TV series *Extrapolations*, an American production broadcast on the Apple TV+ platform this year, and created by Scott Z. Burns. It is an eight-episode miniseries that depicts a relatively near future in which the climatic effects that the scientific world has been predicting for more than twenty-five years are beginning to be felt. The first story takes place in 2037 and the last in 2070. The aim of our work will be to analyse the different ethical questions raised in the series and the relationship they have with the configuration of today's Information Society. To do so, we will analyse the issues that circulate in the press and see how they are reflected in the series, , with an integral perspective from an eco-feminist perspective.

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 Research Incentive for Young PhDs, in the context of the V PRICIT (Regional Programme of Research and Technological Innovation). Call PR/27/21. Title: "Traceability, Transparency and Access to Information: Study and Analysis of the dynamics and trends in the area". Reference: PR27/21-017. Duration: September 2022 - December 2024. Funding of 43,744.22 euros.

KEYWORDS: Mainstream cinema, climate crisis, Ethics, Social Change, Eco-feminism.

REFERENCES

Bernárdez Rodal, A. (2023). *Ecoficciones. Cine para sentipensar la crisis climática.* Valencia: Tirant lo Blanch.

Bruna Pérez, P. (2020) Análisis del imaginario cinematográfico de posibles futuros ecosociales y alternativas narrativas especulativas. *Re-visiones*, (10), 1-22.

Canavan, G., & Rovinson, K. S. (Eds.) (2019). *Green Plantes: Ecolgy and science fiction*. Middletown CT: Wesleyan University Press.

Demos, T. J. (2017). *Against the Anthropocene. Visual culture and environment today.* Berlin: Stenberg Press.

Dederichs, N. (2023). *Atmosfears: The Uncanny Climate of Contemporarry Ecofiction.* Transcript Verlag.

Hameed, A.; Gunkel, H., & O'Sullivan, S. D (Eds.) (2022). *Futures and Fictions*. London: Reapeater.

Jameson, F. (2005). *Archaeologies of the Future: The Desire Called Utopian and Other Science Fiction*. London & New York: Verso.

Montoro Araque, M. (2023). ¿Hacia un fantástico ecoficcional? Dos lecturas de lo monstruoso vegetal en el cine contemporáneo. *Brumal. Research Journal on the fantastic, 11*(1), 211-229. https://doi.org/10.5565/rev/brumal.959

Morto, T. (2016). *Dark Ecology. For a Logic of future coexistence*. New York: Columbia UniversityPress.

WOMEN COMMUNICATORS IN TIKTOK. KEYS FROM THE TRACEABILITY OF THE INFORMATION WITH A GENDER PERSPECTIVE

Graciela Padilla-Castillo, Asunción Bernárdez-Rodal, Ignacio Moreno-Segarra, Jonattan Rodríguez-Hernández

Complutense University of Madrid (Spain)

gracielp@ucm.es; asbernar@ucm.es; igmore01@ucm.es; jonrodri@ucm.es

EXTENDED ABSTRACT

The main studies on social media agree that the time spent on them is increasing year on year compared to the time spent on traditional media (Ortega, Padilla & Vaquerizo, 2021; Padilla & Rodríguez, 2022; Rodríguez Hernández, 2022a). This change refers to audiences of many age strata, not just the alpha, centennial and millennial generations (El Habchi & Padilla, 2020; Ortega & Rodríguez, 2021; Padilla Castillo, 2023). On the other hand, the change is seen as negative, as if the information on networks were of poorer quality, more biased and with more hoaxes and fake news (Bernárdez, López & Padilla, 2021; Rodríguez Hernández, 2021; Requeijo, Padilla & Díaz, 2022; Rodríguez Hernández, 2022b). However, this proposal opts for an objective exploration, without previous negative or positive hypotheses, to study the possible paradigm shift and the characteristics and circumstances of this information on networks with a gender perspective. Specifically, it focuses on an analysis of TikTok and the 1-minute news programmes in Spanish, which have become one of the most successful formats on the Chinese social network. In them, women communicators offer short, 60-second news programmes, summarising current affairs for their audience. Through the results of the study, this paper argues the importance of gender's perspective is even more important to keep in mind when the possible change of paradigm of audiences in TikTok and other social media.

TikTok is a social networking platform that focuses mainly on the creation and sharing of short videos (1, 3 or 5 minutes). Although TikTok users include people from different professions and fields, communication professionals still seem reluctant to appear on this social network. The same is true for many companies and official institutions, which do not want to open an account or which open an account, often receiving a lot of criticism from the audience, which wrongly associates TikTok with a lack of seriousness (Ortega & Rodríguez, 2021; Rodríguez Hernández, 2022c; Padilla Castillo, 2023). However, we believe that this social network should be studied as the mass communication phenomenon that it is; and despite its errors or possible ethical problems, its audience and engagement data make such an analysis necessary. Even more so when the latest global reports continue to point to the growth of young and adult users (alpha, millennial and centennial generations), and how they choose the application over traditional media to get information on current affairs in general or on specific topics. As many academics and practitioners say, TikTok is not a social network of people dancing.

This work is part of a coordinated project between Spain, Portugal and the United Kingdom, whose main objective is to study the traceability of information in order to combat disinformation among citizens. European bodies have set out different initiatives to combat disinformation and promote free access to information. However, the recommendations do not always become obligations, they are very varied and sometimes local, and not enough

improvements have been made to improve journalistic dynamics and citizens' knowledge of public information. At the same time, social networks and the appearance of news on media accounts, together with premium subscriptions to digital newspapers, have made the situation more complex. In these circumstances, it is understandable that the 1-minute news programmes on TikTok are multiplying in versions and number of viewers, as they stand as a fast, convenient alternative, adapted to each person's schedule and habits.

To understand the possible change of paradigm and this new infotainment format, a mixed methodology is used: quantitative exploration of accounts and audiences in TikTok of the most successful news programmes in Spanish with women communicators; qualitative analysis of styles, video formats, use of infographics and emojis, presenters' styles and topics covered with a gender perspective. The field study covers the 30 female tiktokers that summarise, in Spanish, the daily news in 1-minute news programmes, with the quantitative and qualitative items described above.

The results show higher audiences than Kantar Media and EGM data for Spanish news programmes; high audience engagement in terms of interactions and comments; and a surprising coincidence of topics between 1-minute news programmes and, at the same time, between TikTok news programmes and traditional media news programmes. Traditional media news anchors often have specific and traditionally established roles, reporting on current events, politics, entertainment, sports and other relevant topics. These professionals are trained in telegenic and possess specific skills in journalism, effective communication and on-camera verbal and non-verbal communication.

However, the female communicators of these new TikTok news programmes do not have the same training and their audience, in some cases, is in the millions. Among the keys to their success, we can find several possibilities: they offer informative and relevant content for their community; they are concise and direct, with short and impactful messages; they visually support their words with filters and visually striking backgrounds; they develop more varied and natural body languages, with different gestures and postures compared to traditional news programmes; they emphasise certain news by playing with their tone of voice; they bring their personality and opinion to the news narrative; they manage to create a sense of closeness or familiarity with the user; they employ humour and make the news seem even enjoyable.

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 Research Incentive for Young PhDs, in the context of the V PRICIT (Regional Programme of Research and Technological Innovation). Call PR/27/21. Title: "Traceability, Transparency and Access to Information: Study and Analysis of the dynamics and trends in the area". Reference: PR27/21-017. Duration: September 2022 - December 2024. Funding of 43,744.22 euros.

KEYWORDS: TikTok, News programmes, Information, Traceability, Female communicators.

REFERENCES

Bernárdez-Rodal, A., López-Priego, N., & Padilla-Castillo, G. (2021). Culture and social mobilisation against sexual violence via Twitter: the case of the “#LaManada” court ruling in Spain. *Revista Latina de Comunicación Social*, *79*, 237-262. https://www.doi.org/10.4185/RLCS-2021-1502

El Habchi Mahir, Z. & Padilla Castillo, G. (2020). Leadership and Authority Communication In Social Networks. The Case of Lady Amelia Windsor in Instagram. *Revista Internacional de Investigación en Comunicación aDResearch ESIC*, *23*(23), 70-89. https://doi.org/10.7263/adresic-023-04

Padilla Castillo, G. (2023). "The Melancholy Bubble". Emotional dangers in social media. *Human Review. International Humanities Review*, *16*(6), 1-10. https://doi.org/10.37467/revhuman.v12.4703

Padilla Castillo G. & Rodríguez Hernández J. (2022). Sustainability in TikTok after COVID-19. The viral influencers in Spanish and their micro-actions. *Estudios sobre el Mensaje Periodístico*, *28*(3), 573-585. https://doi.org/10.5209/esmp.81133

Ortega Fernández, E., Padilla Castillo, G, & Vaquerizo Domínguez, E. (2021). Píldoras audiovisuales y enseñanza universitaria en Comunicación. Ruptura de la brecha digital y nuevas competencias. *Bibliotecas. Anales de Investigación*, *17*(4), 1-19. Retrieved from: http://revistas.bnjm.cu/index.php/BAI/article/view/451

Ortega Fernández, E., & Rodríguez Hernández, J. (2021). Communication Strategy of Security Forces through Audiovisual Pills In TikTokNational Police and Civil Guard in Spain. *Revista Internacional de Investigación en Comunicación aDResearch ESIC*, *25*(25), 160-185. https://doi.org/10.7263/adresic-025-09

Ortega Fernández, E., y Rodríguez Hernández, J. (2021). Hashtags empoderadores en favor de todos los tipos de cuerpos: #BodyPositive en Instagram y TikTok. In A. Bernárdez & G. Padilla (Eds.), *Deshaciendo nudos en el Social Media* (pp. 343-364). Valencia: Tirant Lo Blanch.

Requeijo Rey, P., Padilla Castillo, G., & Díaz Altozano, P. (2022). Transphobia on Twitter. The Rachel Levine Case at the Start of Joe Biden's Presidency. *Fonseca, Journal of Communication*, (25), 181-204. https://doi.org/10.14201/fjc.29742

Rodríguez Hernández, J. (2021). Uso de Twitter en el periodo postelectoral estadounidense Donald Trump y Joe Biden. In B. Sánchez-Gutiérrez & A. Pineda (Eds.), *Comunicación política en el mundo digital: tendencias actuales en propaganda, ideología y sociedad* (pp. 447-479). Madrid: Dykinson.

Rodríguez Hernández, J. (2022). La guerra en TikTok. La red social de la invasión rusa a Ucrania. In L. R. Romero-Domínguez & N. Sánchez-Gey Valenzuela (Eds.), *Sociedad digital, comunicación y conocimiento. Retos para la ciudadanía en un mundo global* (pp. 150-168). Madrid: Dykinson.

Rodríguez Hernández, J. (2022). Información en tiempos de TikTok. Medios de comunicación que superan el millón de seguidores en la red social china. In G. Paredes Otero (Coord.), *Narrativas y usuarios de la sociedad transmedia* (pp. 818-834). Madrid: Dykinson.

Rodríguez Hernández, J. (2022). Museums and TikTok: Bringing Art to Young People. *VISUAL REVIEW. International Visual Culture Review / Revista Internacional De Cultura Visual*, *11*(3), 1-10. https://doi.org/10.37467/revvisual.v9.3677

5. Smart Education: transforming learning in the digital age

Mario Arias-Oliva, Complutense University of Madrid; Antonio Pérez-Portabella, Universitat Rovira i Virgili; Teresa Pintado, Complutense University of Madrid; Joaquín Sánchez Herrera, Complutense University of Madrid

EXAMINING THE PARALLEL MEDIATING EFFECT OF FINANCIAL EDUCATION AND CORPORATE ETHICS ON THE RELATIONSHIP BETWEEN FINANCIAL FRAUD RISK AND INTENTION TO USE FINANCIAL SERVICES: IMPLICATIONS FOR UNIVERSITY CURRICULA

Pedro I. González-Ramírez, Juan Carlos Yáñez-Luna

Universidad Autónoma de San Luis Potosí (México)

pedro.gonzalez@uaslp.mx; jcyl@uaslp.mx

EXTENDED ABSTRACT

Fraud poses a significant challenge to the global economy. Its negative effects extend from the stability of financial markets to businesses, consumers, and governments. Confronting this large-scale issue requires a comprehensive approach that involves multiple stakeholders. In this regard, one of the primary reasons international financial institutions are concerned with combating financial fraud is to preserve the integrity of the global financial system. Fraud can erode investor and market participants' trust, leading to capital flight, financial volatility, and decreased investment. Moreover, fraud can hinder access to credit and basic financial services, thus impeding economic and social development. These consequences can undermine economic growth and financial stability, impacting entire countries and regions.

The term "fraud" can be simply defined as a crime involving deceptive activities in a specific sector. These activities are typically carried out by individuals or groups with malicious intent to obtain illicit benefits. With the increased use of technology in recent decades and the interconnectedness of devices (Internet of Things), such illegal activities have become more prevalent. For instance, Cross et al. (2014) note that online fraud arises from "an individual's experience of responding over the internet to a dishonest invitation, solicitation, notification, or offer, by providing personal information or money that results in a loss, with or without financial impact." Likewise, Juhandi et al. (2020) state that fraud refers to "a broad legal concept, describing any intentional fraudulent attempt aimed at taking someone's property or rights, or those of other parties." Marabad (2021) suggests that "fraud is defined as an unlawful deception intended to secure financial or personal gain. It is a premeditated behavior that goes against the law or policy to achieve unfair financial gain".

The abovementioned concepts propose cooperation among local and international financial institutions, government authorities, and regulatory bodies to create public policies that effectively combat financial fraud in a region (Zamudio et al., 2022). For instance, the OECD highlights that implementing effective programs to promote financial inclusion and education can enhance consumer awareness, which, in turn, are key elements for protecting against and preventing fraud while fostering a culture of integrity and transparency in the financial sector. In this regard, the economic impact of financial fraud in emerging countries can have significant consequences. This means that financial institutions and consumers bear direct financial losses while the overall economy suffers from a lack of trust and decreased investment. Moreover, financial fraud can erode the reputation of companies and the country, leading to long-term effects on economic development.

Given the above, the government in Mexico has implemented several measures to combat financial fraud, including the establishment of specialized units for financial crimes and the enactment of stricter laws and regulations (CONDUSEF, 2021). However, it is worth noting that the country still faces significant challenges related to infrastructure, such as telecommunications and cybersecurity. In this regard, the lack of robust security infrastructure can impact financial institutions' reputation and perceived quality (García Witron, 2021), ultimately affecting customer loyalty. Building trust remains a recurring challenge for institutions, particularly those operating in the financial sector. Institutions must develop a corporate ethics framework as a social normative framework, embracing ethical practices to convey commitment and social responsibility as a loyalty-building strategy (Gómez Pescador & Arzadun, 2019).

This study assesses the relationship between financial fraud risk, corporate ethics, and the intention to use financial services through a financial education framework in the university curriculum. Notably, the study will evaluate the direct effect between financial fraud risk and the intention to use financial services. Furthermore, the study proposes to assess the mediating impact of corporate ethics on the direct relationship between financial fraud risk and the discretion to use financial services, as well as the mediating effect of financial education on the direct relationship between financial fraud risk and the intention to use financial services.

The methodological justification of this study is based on the need to comprehensively address the complex relationships and potential underlying mechanisms between the investigated variables: financial education, corporate ethics, financial fraud risk, and intention to use financial services. This study adopts a methodology based on PLS-SEM (Partial et al. Equation Modeling) and path analysis (Hair et al., 2013) to better understand how these variables interrelate and how they may influence the intention to use financial services. Applying PLS-SEM and path analysis enables us to take an integrated approach by considering all the mentioned variables and their interactions within a single model. This methodological approach provides a more precise and holistic understanding of the complexities and interrelationships among the variables of interest.

According to the proposed methodology, the following relationships can be examined using PLS-SEM:

1. Relationship between Financial Education and Intention to Use Financial Services:
    - Hypothesis: Financial education positively influences the intention to use financial services.
    - Path: Financial Education $\rightarrow$ Intention to Use Financial Services.
2. Relationship between Corporate Ethics and Intention to Use Financial Services:
    - Hypothesis: Corporate ethics positively influence the intention to use financial services.
    - Path: Corporate Ethics $\rightarrow$ Intention to Use Financial Services.
3. Relationship between Financial Fraud Risk and Intention to Use Financial Services:
    - Hypothesis: Financial fraud risk negatively influences the intention to use financial services.
    - Path: Financial Fraud Risk $\rightarrow$ Intention to Use Financial Services.
4. Mediating Effect of Corporate Ethics on the Relationship between Financial Fraud Risk and Intention to Use Financial Services:

- Hypothesis: Corporate ethics mediates the relationship between financial fraud risk and the intention to use financial services.
- Paths: Financial Fraud Risk → Corporate Ethics → Intention to Use Financial Services.

5. Mediating Effect of Financial Education on the Relationship between Financial Fraud Risk and Intention to Use Financial Services:

   - Hypothesis: Financial education mediates the relationship between financial fraud risk and the intention to use financial services.
   - Paths: Financial Fraud Risk → Financial Education → Intention to Use Financial Services.

By measuring previous relationships and mediating effects through PLS-SEM, the study can provide insights into the direct and indirect influences of financial education, corporate ethics, and financial fraud risk on the intention to use financial services. This analytical approach allows for a comprehensive understanding of the underlying mechanisms and the potential parallel mediating effects among these variables in the study context.

The aim is to obtain results that enable decision-makers to formulate strategies and policies based on corporate ethics, creating a context of trust in using financial services even in the presence of financial fraud risk. The proposed model will also demonstrate a positive relationship between incorporating financial education in university curricula and the intention to use financial services. Thus, by establishing a reliable and ethical environment, financial institutions can strengthen their relationship with clients and encourage greater engagement in financial services, thereby contributing to economic and social development.

The findings in this study will have implications for the design of university curricula in finance or related disciplines. Courses or modules focusing on financial education and corporate ethics can equip students with the necessary knowledge and ethical principles to navigate the financial landscape effectively. By addressing the mediating effects of financial education and corporate ethics, universities can contribute to developing professionals who are both knowledgeable and ethically conscious, promoting a safer and more trustworthy financial services ecosystem.

KEYWORDS: Financial Education, University Curricula, Corporate Ethics, Financial Fraud Risk, Intention to Use Financial Services.

REFERENCES

CONDUSEF. (2021). Fraudes cibernéticos y tradicionales. Secretaría de Hacienda y Crédito Público. https://www.condusef.gob.mx/documentos/comercio/FraudesCiber-3erTrim2019.pdf

García Witron, C. (2021). Ciberseguridad en el Sector Financiero. ¿Cómo transformar una amenaza en una oportunidad? Trabajo Fin de Grado. Universidad Pontificia de Comillas. http://hdl.handle.net/11531/46570

Gómez Pescador, I., & Arzadun, P. (2019). Responsabilidad social cooperativa en el sector de ahorro y crédito de costa rica. Mediación de la reputación, credibilidad y percepción en la lealtad de los asociados. *Boletín de Estudios Económicos*, 74(228), 553–578.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2013). Partial Least Squares Structural Equation Modeling: Rigorous Applications, Better Results and Higher Acceptance. Long Range Planning, 46(1-2), 1-12. https://doi.org/10.1016/j.lrp.2013.01.001

Juhandi, N., Zuhri, S., Fahlevi, M., Noviantoro, R., Nur abdi, M., & Setiadi. (2020). Information Technology and Corporate Governance in Fraud Prevention. E3S Web of Conferences, 202, 16003. https://doi.org/10.1051/e3sconf/202020216003

Marabad, S. (2021). Credit Card Fraud Detection using Machine Learning. Asian Journal of Convergence in Technology, 7(2), 121–127. https://doi.org/10.33130/AJCT.2021v07i02.023

Zamudio, L. F., Sauceda, A. L., & Ramos, B. A. (2022). Educación financiera para nivel de educación media superior: caso CECYTE, Baja California, México. Espacios, 43(11), 13–24. https://doi.org/10.48082/espacios-a22v43n11p02

CYBERSECURITY EXPERIENTIAL LEARNING EDUCATION

Shalini Kesar, Ashely Tyler

Southern Utah University, (USA)

Kesar@suu.edu; ashleytyler@suu.edu

EXTENDED ABSTRACT

This paper is part of the collaborative on-going research between the author and co-author (PI of an educational grant in the Forest Service division). The initial research began a few years ago with the idea of developing or modifying the design curriculum to provide an educational experiential learning using the right ethical practices of National Society for Experiential Education (NSEE, 2009) framework (Kesar and Pollard, 2020, 2021). It started with focusing on undergraduate students in STEM field (computer science, information systems, cybersecurity and technology). As the research has progressed, the context has shifted to online graduate students in cybersecurity. This paper sheds light on how author collaborated with the co-author to design the class and how it was beneficial in creating team building project in cybersecurity as well as adding value to the collaborator.

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 the knowledge, skills and attitudes that empower learners and promote the common good (NSEE, 2023). The framework consists of eight principals linked with good practices. In this paper, the project conducted with graduate cybersecurity students is discussed that was designed by the instructor (author) as part of an experiential learning activity. The goal was that this experience and the learning will add value to the fundamental of creating an online cybersecurity training as part of a group project. While the authors (instructor and client) collaborated and designed the training, it is hoped that all the parties are empowered to use the right principals mentioned in the framework. Consequently, ensuring both the quality of the learning experience and of the work produced by the students, and in building an assignment that underlie the pedagogy of experiential education. Although the NSEE framework was used, the main thought process was very different when designing the project. It considered the framework as well the research regarding the team projects and importance 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. While developing the team project curriculum, various studies were taken into account, including author's previous published research. Best standards and Guiding Principles of Ethical Practice by the National Society for Experiential Education (NSEE) were used. The NSEE Guiding Principles of Ethical Practices are used to develop the pedagogy to teach ethics and professional as part of an experiential education. This paper describes how the how instructor included ethics and professionalism in this team project. The eight principals are exampled below.

Intention: In this principal, all parties must outline a clear vision on the reason which this particular experience is chose and the why experience is the chosen approach to the learning. It is expected that the assignment linked with cybersecurity training for the online graduate students is to allow them to share knowledge with their team members as well as demonstrate, apply or result from it. The principal of Intention in general 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.

Preparedness and Planning: The main objective of this principal was to ensure the students have a group project experience and each member of the project has a successful experience from the earliest stages of the experience/program. This aligned with the identified intentions, adhering to them as goals, objectives and activities designed as part of the project. As mentioned earlier, the project was to design a cybersecurity training for employees, who are part of the Forest Service division. The project involved students and provided them flexible enough to allow for adaptations as the experience of creating training planning unfolds.

Authenticity: In this principal it is important the students have an experience that is in a real-world context. In this project, the students developed a training program for a small set of employees of the Forest service with the intent the training useful and meaningful as part of the employees' annual required training. The three groups compromising of three to four members developed training programs on different cybersecurity topics including Phishing, Social Engineering, and Passwords.

Reflection: NSEE refers to Reflection as an element that transforms simple experience to a learning experience. With this principal in mind, the assignment was designed so that knowledge can be discovered and internalized as the students 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 weigh as well as reflect the outcomes against past learning and future implications. This reflective process in the assignment comprised 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.

Orientation and Training: The students were required to discuss and show the training hey had developed to the client. This not prepared them work as a team but also experience and learn about each other and about the context and environment in which the training will be presented to the small division of the Forest Services.

Monitoring and Continuous Improvement: In this principal, it is important to note that any learning activity designed should be dynamic and changing. In addition, the instructor (author) outlined the assignment with the student learning outcomes that included reports and presentation that 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 the 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, instructor and client's experience were noted and included in the lessons learned and reflection in the recognition of progress and accomplishment. Given that this was part of an 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.

KEYWORDS: NSEE, Cybersecurity, Training, Pedagogy, online graduate class**.**

REFERENCES

Kesar, S., and Pollard, J. (2021) "Cultivating an Empathic Learning Pedagogy: Experiential Project Management", in N*ormal 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

Kesar, S., and Pollard, J, "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.

The National Society (2009). Guiding Principals of Ethical Practices. https://www.nih.gov/health-information/nih-clinical-research-trials-you/guiding-principles-ethical-research

Jervis, K. J., and Hartley, C. A. (2005). Learning to design and teach an accounting capstone. *Issues in Accounting Education*, 20 (4), 311-339. https://doi.org/10.2308/iace.2005.20.4.311

USE AND ABUSE OF AI – ETHICAL PERSPECTIVES IN THE EDUCATIONAL SECTOR

Isabel Alvarez, Nuno Silva

ISTEC, COMEGI (Portugal), Lusíada University, COMEGI (Portugal)

alvarez@edu.ulusiada.pt; nsas@lis.ulusiada.pt

EXTENDED ABSTRACT

This paper discusses the impact and the potential implications of the generative Artificial Intelligence language model, namely ChatGPT in higher education. Some educators think that this popular bot can alter teaching while others worry that it may have the opposite effect on their students' motivation to learn. Others believe students may benefit from understanding the ins and outs of how this technology works and might use it as a tool to explore the possibilities and limits of online sources of information. Though, apparently, there are benefits of this new technology, a lot of caution is required for its use.

Several publishers have recently introduced new policies in response to the growing use of Generative AI (Artificial Intelligence) applications. It can generate detailed responses to questions related to several subjects hardly distinguishable from those created by humans, which on one side is impressive but on the other side this potential is also very concerning and worrying that could lead to serious problems in education (Yang et al., 2021). Moreover, these technologies enhance learners' abilities in memorizing, comprehending, applying, analysing, and assessing, with the utmost educational objective to the highest cognitive level, which is creativity (Hwang & Chen, 2023).

ChatGPT can play the role of a debate opponent and generate counterarguments by exposing students to an endless supply of opposing viewpoints, helping them to look for weak points in their own thinking (Will, 2023). Beverly Park Woolf research (Woolf et al, 2013) focuses on the use of AI in education, with a particular emphasis on intelligent tutoring systems. Probably the best way will be to schools to start encouraging students critical thinking about what technology can help and what it hinders us from doing instead of just teaching how to use technology (Woolf et al, 2013).

Lecturers are considering using the ChatGPT to plan lessons, offer students feedback on assignments, and execute some administrative tasks. But the technology of ChatGPT is not yet fool proof. Some lecturers published that they noticed a factual error when they experimented asking the bot to plan a lesson for an early chapter on a certain subject. The tool also demonstrated that has limited knowledge of world events that happened after 2021(Will, 2023). ChatGPT can also offer feedback on student work. Other situations have also been reported by lecturers, saying that the examples of grading from the chat bot feel shallow or even inaccurate. It was also published that, while the technology might get it right nine times out of 10, there's always the risk that it won't grade one student's work correctly, so lecturers would still need to personally review each piece of feedback (Will, 2023). Some schools worldwide have decided to ban the use of this bot and issued statements that warned students against using ChatGPT to cheat. And as some authors say, while the tool may be able to provide quick and easy answers

to questions, it does not build critical-thinking and problem-solving skills, which are essential for academic and lifelong success (Will, 2023).

The use of AI in higher education presents both good opportunities and also challenges that need to be addressed by taking a proactive and ethical approach to the use of AI in education (Cotton et al., 2023), namely if it is genuinely useful in supporting teaching and learning (Kousa & Niemi, 2023).

There are some opinions that the threats to education in this context is based on a lack of deep understanding and difficulty in evaluating the quality of responses, threatening academic integrity, democratising plagiarism and declining high-order cognitive skills (Farrokhnia et al., 2023).

While there is much generic literature on ethics in artificial intelligence, there is a clear gap in studies on the ethics of ChatGPT in the education sector. We systemically explore what exists and address what does not exist. For Pedró et al. (2019), the major challenges are related to personalisation, inclusion and equity, powered education, quality, and transparency. The issues of equity and personalisation are detailed by Chine et al. (2022), namely in the case of experience learning gaps due to a lack of access or economic disadvantages. On the other hand, Jiang and Pardos (2021), gives special attention to fairness and bias in artificial intelligence and graduation prediction. Regarding specifically ChatGPT (Cotton et al., 2023), it opens new difficulties of detecting and preventing academic dishonesty. An update of plagiarism detection tools and controlling cheat proctoring tools is absolute necessary. The output from ChatGPT not include proper referencing, while academic writing is expected to accurately include citations and references. The ChatGPT has raised security and privacy issues, namely because there is no minimum age requirement to use ChatGPT. Also, it is not clear that personal data analysis is done in respect to GPDR (EU General Data Protection Regulation).

Even more, there are emerging smart small wearable devices like smartwatches and hearables. How should educators respond when problems like these inevitably occur? (Krutka, Pleasants & Nichols, 2023). Like all technologies, smart digital devices bring unintended, collateral, and disproportionate effects.

The methodological approach is mainly to engage in reflective practice concerning the adoption and use of artificial intelligence in higher education, and authors as lecturers are exploring the case of ChatGPT in tree Portuguese Universities contexts. The methods used are based on our daily experience observing students and institutions academic activities, qualitative interviews and discussion boards.

In conclusion, the use of ChatGPT in education can bring many potential benefits, such as personalized learning, better feedback, and enhanced student engagement. However, it is important to use in a responsible and ethical manner that respects the privacy and well-being of students, as well as the principles of good teaching. ChatGPT is not designed to address issues related to accountability and cybersecurity directly. The alarm generated by news and evidence reported on the potential of ChatGPT forced an ethical reflection in practical context that the authors as lecturers framed in the education sector. The use and abuse of ChatGPT is not yet verified in the classroom environment. There are no regulatory recommendations or guidelines. The control on plagiarism in autonomous work makes it necessary to reinforce the oral assessment. The model itself is not copyrighted, but the content generated by ChatGPT may be subject to copyright laws. In addition, there are problems related to equity and autonomy granted to students, especially the ability that ChatGPT gives them to do practical work in an assertive way and in a short period of time. Plagiarism

is a major concern, and this involves not only the work of the lecturer but also academic regulations. While AI language models cannot avoid plagiarism on their own, students should take steps to ensure that any content produced using these tools is properly cited and attributed to its original source. Ultimately, the extent to which ChatGPT is aligned with educational goals and values will depend on how the model is used and the degree of care taken to ensure that its responses are accurate, relevant, and appropriate for the educational context in question.

KEYWORDS: Artificial Intelligence, Ethics, ChatGPT, Higher Education.

ACKNOWLEDGEMENTS: This work is supported by national funding's of FCT - Fundação para a Ciência e a Tecnologia, I.P., in the project «UIDB/04005/2020».

REFERENCES

Chine, D., Brentley, C., Thomas-Browne, C., Richey, J., Gul, A., Carvalho, P., Branstetter, L., & Koedinger, K. (2022). Educational equity through combined human-AI personalization: A propensity matching evaluation. In *International Conference on Artificial Intelligence in Education* (pp. 366-377). Springer, Cham.

Cotton, D., Cotton, P., & Shipway, J. (2023). Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT. *Innovations in Education & Teaching International*. https://doi.org/10.1080/14703297.2023.2190148

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT Analysis of ChatGPT: Implication for Educational Practice and Research. *Innovations in Education & Teaching International*. https://doi.org/10.1080/14703297.2023.2195846

Hwang, G.-J., & Chen, N.-S. (2023). Editorial Position Paper: Exploring the Potential of Generative Artificial Intelligence in Education: Applications, Challenges, and Future Research Directions. *Educational Technology & Society*, *26*(2). https://doi.org/10.30191/ETS.202304_26(2).0014

Jiang, W., & Pardos, Z. A. (2021). Towards Equity and Algorithmic Fairness in Student Grade Prediction. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)* (pp. 608-617). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/ 3461702.3462623

Kousa, P., & Niemi, H. (2023). Artificial Intelligence Ethics from the Perspective of Educational Technology Companies and Schools. In H. Niemi, R. D. Pea, & Y. Lu (Eds.), *AI in Learning: Designing the Future*. Springer. https://doi.org/10.1007/978-3-031-09687-7_17

Krutka, D. G., Pleasants, J., & Nichols, T. P. (2023). Talking the Technology Talk. *Phi Delta Kappan*, *104*(7), 42-46.

Pedró, F., Subosa, M., Rivas, A., & Valverde, P. (2019). *Artificial intelligence in education: challenges and opportunities for sustainable development*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000366994

Will, M. (2023, January 11). With ChatGPT, Teachers Can Plan Lessons, Write Emails, and More. What's the Catch? *Education Week*. https://www.edweek.org/technology/with-chatgpt-teachers-can-plan-lessons-write-emails-and-more-whats-the-catch/2023/01

Woolf, B. P., Lane, H. C., Chaudhri, V. K., & Kolodner, J. L. (2013). AI Grand Challenges for Education. *AI Magazine*, *34*(4), 66-84. https://doi.org/10.1609/aimag.v34i4.2490

Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, *2*, 100008. https://doi.org/10.1016/j.caeai.2021.100008

CHAT GPT: HAS ITS POTENTIAL ARRIVED TO ENHANCE THE NEW WAY OF TEACHING AND LEARNING? A CASE STUDY IN AVIATION STUDIES

Ana María Lara Palma, Rafael Brotóns Cano

Universidad de Burgos (Spain), Antolín S.A. (Spain)

amlara@ubu.es; rafael.brotons@antolin.com

EXTENDED ABSTRACT

Teaching and learning are two concepts that are intrinsically linked. The excellence of the latter depends on the innovation of the former. Therefore, the resources that teachers use to update each discipline are the guiding thread towards quality didactics. And, in this line of innovation, are the digital tools. The basic mandates and lines of action in the field of education established in the European Higher Education Area indicate that, "it is the responsibility of the universities to ensure that the studies are innovative and original, incorporate lines that favour the development of the professional career, take into account the importance of inclusion and diversity, serve for social use and the consequent academic support is given to achieve excellence" (European University Association). Additionally, The European Commission in its Digital Education Action Plan (2021-2027) sets the objective of readjusting education and training to the digital age. It highlights two main guidelines. Firstly, to promote the development of a high-performing digital education ecosystem and, secondly, to enhance digital skills and competences for the digital transformation (Lara-Palma, 2022).

Artificial intelligence AI in higher education has contributed as a supporting element whose benefits have been indisputable so far. But everything evolves, and computer systems have been adapting to an increasingly faster market committed to user satisfaction. ChatGPT, a repository of content generated by artificial intelligence has arrived as an assistant to higher education where interaction is done through a chatbot which provides detailed and precised answers (Kocón, 2023); furthermore, adding a challenging language and generation tasks in the form of conversation (Wu, T. 2023).

This new scenario opens the possibility of analysing the benefits and limitations of its use in the classroom from a bidirectional perspective, that of the teacher and that of the student. Therefore, the aim of this study is to analyse the ChatGPT tool usability in the lectures by with the following question: is ChatGPT a resource that reinforces acquisition of learning in the classroom?

Regarding the methodology, two surveys were developed (one for teachers and one for students), which were conducted in paper form (in the winter semester of the academic course 2023-2024). They consisted in 10 questions with special emphasis in the relevance of ChatGPT as learning resource (for students) and as teaching resource (for teachers). All questions are answered by assigning a score number ranging from 1 (leftmost option) to 5 (rightmost option).

Both, students and teachers completed the surveys in an anonymous way in order to prevent unintended data recollection and to encourage all of them to answer in the most honest way possible. The total sample is composed by 22 students (from Commercial Pilot for Passenger and

Cargo Transport Degree at Burgos University) and 5 proffessors (currently working at academic institutions and private companies).

A reduce sample of the student´s questionnaire is included in Table 1. Each question has been assimilated to a representative boundary/drawback and additionally we tested a sentiment analysis (like emotion recognition).

Table 1. Student's Survey.

Boundaries	**Drawbacks**
ChatGPT is always available	Plagiarism must be taken into account
ChatGPT is entertaining	Doubts about authenticity
ChatGPT is free of charge	It is not well seen
ChatGPT is the same as talking with a proffessor	No emotion recognition
ChatGPT is a Master of all disciplines	I no longer read books or articles
ChatGPT is fast. I do not wate my time	I can ask in English or any other language
ChatGPT is out of class support	There are no figures or tables
ChatGPT fits my needs perfectly	I don't need to attend tutorials
ChatGPT is easy to access	Less interaction with my classmates
ChatGPT is an intelligent turoring system	Less independent

Source: Self-elaboration and based on Fawaz (2023)

The obtain results provide a basis for a fundamental discussion of whether ChatGPT is a useful digital resource that can enhance learning (for students) and teaching (for teachers). Moreover, can provide customized strategies and approaches to students' characteristics and needs (Crompton, 2023). It has undoubtedly been a cultural impact with multiple implications for cybersecurity and education, among several other disciplines.

As conclusions, the study allows to address the Code of Ethics of use and rules of responsibility (influence of AI models in learning, avoiding plagiarism, not influencing creativity, shortcomings in the teacher-student relationship or, something as essential in the academic work as it is to promote the acquisition of soft skills and disciplinary competences).

KEYWORDS: ChatGPT, learning threat, academic innovation, aviation studies.

REFERENCES

Crompton H., Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20 (22), https://doi.org/10.1186/s41239-023-00392-8

European University Association (2007): Doctoral Programmes in Europe´s Universities: Achievements and challenges. Report prepared for European Universities and Ministers of Higher Education. European University Association Publications. 2007.

European Commission/EACEA: The European Higher Education Area in 2020. Bologna Process Implementation Report. 2020.

Fawaz, Q. (2023). ChatGPT in scientific and academic research: future fears and reassurances. *Library Hi Tech News*. 3, pp. 30-32. Emerald Publishing. http://doi.org/10.1108/LHTN-03-2023-0043

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y. (2023). How close is ChatGPT to Human Experts? Comparison corpus, Evaluation and Detection. Retrieved from https://arxiv.org/abs/2301.07597

Kocón, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydlo, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocón, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, L., Wojtasik, K., Woźniak, S., Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion,* 99. https://doi.org/10.1016/j.inffus.2023.101861

Lara-Palma, A. M., Brotóns Cano, R., Valencia, O., Matsuda, D. (2022). Heading for inter-disciplinary lectures: an international collaborative team activity carried out between an Asian and European Universities. 16TH International Technology, Education and Development Conference. IATED.

Nikolic S., Daniel S., Haque R., Belkina M., Hassan G.M., Grundy S., Lyden S., Neal P., Sandison C. (2023). ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education*. http://doi.org/10.1080/03043797.2023.2213169

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q., Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potencial future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5), pp.1122-1136. http://doi.org/10.1109/JAS.2023.123618

BUILDING GLOBAL AWARENESS AND ETHICAL DECISION-MAKING SKILLS IN U.S. BUSINESS STUDENTS: A CALL FOR TECHNOLOGY BASED EXPERIENTIAL LEARNING

Martha Wilcoxson

Colorado State University Pueblo (United States)

Martha.Wilcoxson@CSUPueblo.edu

EXTENDED ABSTRACT

One aspect of business ethics education that has been a major roadblock for U.S. educators is how to effectively teach business ethics in a global economy. The problem stems from three major hurdles. First is the reliance of U.S. institutions on grounding the subject of business ethics in Western ethos, which does not accommodate other ethical standards (Peppas, 2002). In Western institutions, the ethics curriculum begins with the Western philosophy of ethics, which includes Aristotle, Plato, and Kant (White and Taft, 2004). Educators must work harder to teach competencies in non-Western ethical standards (Stein, 2019). The challenge is to build understanding without compromising personal ethical ideals. Learning platforms that foster dialogue, observation, and consideration of a wide range of ethical standards are critical. The second challenge is educator awareness of global issues and the need for instructors to adopt a global perspective in the classroom. Educators must be globally aware if we are to prepare our business graduates to be able to function ethically in international markets and situations. Ethics in international business is complex and requires curriculum that is current as well as instructors who are informed on current global issues (Leclair, 2000). Business educators must be well informed on diverse cultural perspectives (Miglietti, C., 2015). Much of the angst of instructor teaching and student learning in a global arena begins with how business ethics is taught. Finally, the third challenge is students' lack of enthusiasm for understanding the nuances of cultural differences in ethical decision-making (El Baradei, 2020). Building an understanding of a diverse range of cultural norms that impact business decision-making global stage that intrigues and inspires student interest is not easy (Ortiz, J. (2004). Current research indicates the answer to these three hurdles may be in experiential instruction. Creating opportunities for students to experience and interact with students, instructors, and businessmen and women from around the world is the first step in addressing three major hurdles in global business ethics instruction. Technology that facilitates dialogue and interactions between students and international representatives promotes cultural and society awareness in global business ethics and is critical in preparing global business participants (Sanyal, 2000; Pallab & Kausiki, 2005; Saat, 2014). In addition, education that combines conversations on diverse cultural perspectives and practical case scenarios can effectively support instructor skills. The Western business ethics curriculum has historically relied on case studies for teaching global issues. Case studies provide initial but only partial answers. Understanding our role in building ethical organizations in a global society is difficult when the student audience's goal is to live and work in an insular Western business environment. However, as educators, we have an obligation to develop student understanding, and the first step to begin may be case studies that incorporate non-Western perspectives in the form of genuine dialogue. How to achieve true dialogue on a global scale is the key

consideration. Teaching global business ethics requires an appreciation of different perspectives (Witte, 2010). Instructor and curriculum must be competent in understanding and explaining multiple perspectives to teach students to consider ethical issues in relation to the host culture. Meeting the need for a multinational student perspective and preparing Western students to be global business leaders require good teachers and a valid curriculum (Keida and Englis, 2011). Global interconnectivity in business education calls for instructors and curricula that deliver global understanding while preserving nationalism (Rizvi, 2019). Technology has the potential to unite different worlds. The use of virtual experiences and dialogue in global ethics instruction plays a vital role in preparing students to be ethical global leaders. Real-time conversations between business students from different worlds have the potential to build enthusiasm and understanding. Make no mistake; there are challenges in generating conversations between regions with geographic, language and time differences. However, a curriculum that builds real conversations may be the first step to help students appreciate the nuances of global business (Glass & Bonnici, 1997). This paper addresses the need for Western business schools prepare business graduates to operate effectively in a global business environment. Moving beyond Western business ethics taught in the classroom is fresh territory and raises the question of who will be responsible for designing the rules and guidelines for technology-driven global education. The message is simple: we Western educators need to expand our instructional horizons if we are to prepare business students to be leaders in a sustainable global economy.

KEYWORDS: Business ethics, globalization, ethics instruction, ethics competencies, experiencial learning, global competencies.

REFERENCES

Budden, C. B., & Budden, M. C. (2011). It is a small world after all: Teaching business ethics in a global environment. *American Journal of Business Education* (AJBE), 4 (1). https://doi.org/10.19030/ajbe.v4i1.1276

El Baradei, L. (2021) Ethics education in public affairs programs: What do faculty around the globe have to say? *Journal of Public Affairs Education,* 27:2, 198-217, http://doi.org/10.1080/15236803.2020.1818023

Glass, R.S., Bonnici, J., An experiential approach for teaching business ethics. *Teaching Business Ethics,* 1, 183–195 (1997). https://doi.org/10.1023/A:1009793422982

LeClair, D. T., & Ferrell, L. (2000). Innovation in experiential business ethics training. *Journal of Business Ethics*, 23(3), 313–322. http://www.jstor.org/stable/25074247

Kedia, B., Englis, P., (2011), Transforming business education to produce global managers, Business Horizons, Vol. 54 Issue 4, pp. 325-331

Miglietti, C., (2015) Teaching business classes abroad: How international experience benefits faculty, students, and institutions, *Journal of Teaching in International Business*, 26:1, 46-55, http://doi.org/10.1080/08975930.2014.929513

Ortiz, J. (2004). International business education in a global environment: A conceptual approach. *International Education Journal*, 5, 255-265.

Peppas, S., (2002), Attitudes towards business ethics: where East doesn't meet West", *Cross Cultural Management: An International Journal*, Vol. 9 No. 4, pp. 42-59. https://doi.org/10.1108/13527600210797488

Pallab P, Kausiki M., (2005) Experiential learning in international business education, *Journal of Teaching in International Business,* 16:2, 7-25, http://doi.org/10.1300/J066v16n02_02

Rizvi, F. (2019), Global interconnectivity and its ethical challenges in education. *Asia Pacific Education Review.* 20, 315–326 (2019). https://doi.org/10.1007/s12564-019-09596-y

Saat, M. M. (2014). Using experiential learning in teaching business ethics course. In The Clute Institute International Academic Conference Munich, Germany.

Sanyal, R, (2000), An experiential approach to teaching ethics in international business. *Teaching Business Ethics* 4, 137–149. https://doi.org/10.1023/A:1009826909760

White, J., & Taft, S. (2004). Frameworks for teaching and learning business ethics within the global context: Background of ethical theories. Journal of Management Education, 28(4), 463-477.

Witte, E., (2010), The global awareness curriculum in international business programs: A critical perspective, *Journal of Teaching in International Business*, 21:2, 101-131, http://doi.org/10.1080/08975930.2010.483908

6. Smarter Security- Resilience and Recovery

Shalini Kesar, Southern Utah University; Sabina Szymoniak Czestochowa University of Technology

LEGAL AND TECHNICAL CONSIDERATIONS FOR MEDICAL DATA IN HYBRID DATABASE SYSTEM

Olga Siedlecka-Lamch

Department of Computer Science, Czestochowa University of Technology (Poland)

olga.siedlecka@icis.pcz.pl

EXTENDED ABSTRACT

Collecting, storing, and exchanging medical information is an essential aspect of modern healthcare. Relational database systems are extensively used for managing medical data due to their scalability and ability to store structured information. With the advent of blockchain technology, however, a new perspective on the storage and management of medical data emerges (Azaria et al., 2016; Farouk et al., 2020; Linn et al., 2016; Shahnaz et al., 2019).

This article focuses on a hybrid database model that integrates the benefits of relational data storage with the characteristics of blockchain technology. Particular attention will be paid to the legal facets of medical data and the ensuing technical challenges, such as ensuring the right to be forgotten (Rosen, 2011).

Important consideration must be given to the fact that data stored in a blockchain is, in theory, immutable. Existing legal regulations, such as the "right to be forgotten," continue to pose a challenge for medical system providers, who must guarantee the ability to delete data when necessary. In the remainder of the article, we will discuss techniques and strategies that can be effectively implemented in hybrid medical databases to address this issue.

In addition, we will investigate additional legal issues pertaining to medical data, such as privacy protection, compliance with data protection regulations, and controlled data sharing. In addition, we will investigate the technical aspects of implementing hybrid medical databases that facilitate effective data management and legal compliance.

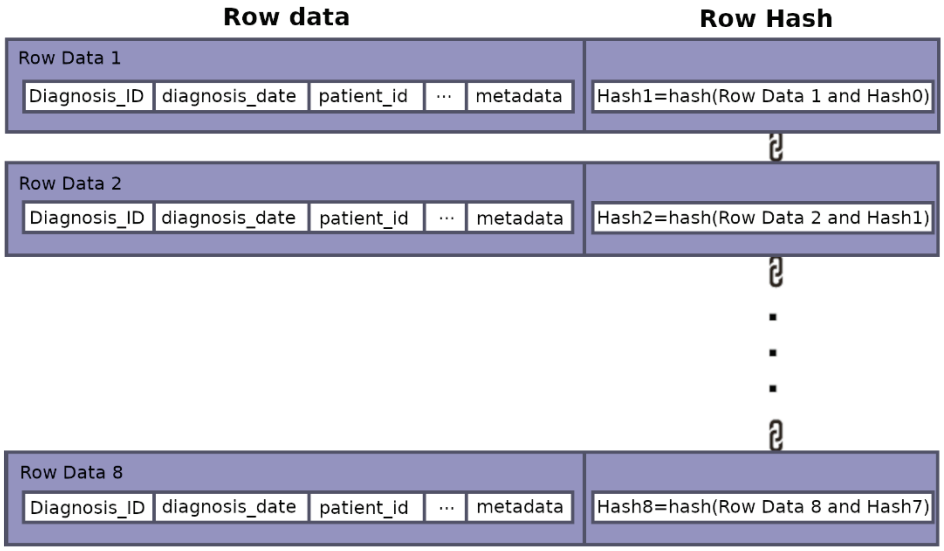
By delving into these issues, this article intends to provide readers with an understanding of the issues surrounding medical databases employing a hybrid model and guidance on the technical solutions that can be utilised to effectively manage medical data and meet legal requirements.

Hybrid model

Numerous researchers have been actively investigating the use of blockchain technology to store medical data for several years. The works of Azaria et al., 2016, Aguiar et al. 2020, Farouk et al., 2020, Linn et al., 2016, Shahnaz et al., 2019, and Yaqoob contain different approaches. Our solution is particularly novel because it enables many medical facilities to leverage existing relational systems by moving sensitive data portions to blockchain tables. This method ensures the immutability of diagnostic and treatment process events without incurring excessive costs associated with transforming entire systems or training staff. The addition of blockchain tables can be incorporated seamlessly, remaining imperceptible to end users and preserving the existing database logic.

We included patient information, medical staff information, visit history, medical leave records, test results, diagnoses, referrals, prescribed medications, disease codes, and their respective categories in our model. The treatment process-related data (test results, diagnoses, prescribed medications) should be stored in a blockchain tables (for example diagnosis table in Figure 1). This will result in an immutable, observable, and easily analysed sequence of events generated by each medical device. It will also be accessible to the patient, but only physicians with the appropriate certificates assigned to their profiles will be able to make changes.

Figure 1. Blockchain table for diagnosis information.



Source: self-elaboration based on Oracle documentation

Obviously, blockchain technology has both benefits and drawbacks, making it difficult to apply it to the entirety of the data (scalability issues, security concerns, certain elements being excessively transparent while others are inaccessible). This is why a combination of technologies can be a highly effective solution, as it combines the advantages of modern innovations with completely functional systems. In the case of our model, achieving legal conformance is the remaining obstacle.

Right to be forgotten

The right to be forgotten is a legal concept that allows individuals to request the removal of their personal information from organisations that acquire and process it. This is especially pertinent in the context of medical data, where the privacy and confidentiality of patient information is essential.

The General Data Protection Regulation (GDPR), which became effective in 2018, has strengthened the right to be forgotten in the European Union. Individuals have the right to request the deletion of their personal data under the GDPR if there are no longer any legal grounds for processing it, if the data is being processed in violation of regulations, or if the individual has revoked their consent for data processing.

Data immutability is the primary characteristic of a blockchain, which means that once transactions are added, they cannot be expunged or altered. In the context of the right to be forgotten and the erasure of medical records, there are a number of methods to overcome this obstacle. Here are some strategies to consider:

- Medical data can be stored off-chain, such as in external file systems or databases, while only the hashes or references to that data are stored in the blockchain. Thus, when the need to expunge the data arises, only the references in the blockchain can be updated or removed without compromising the blockchain's integrity.
- Smart contracts and special functions: Certain blockchains allow for the creation of smart contracts and special functions that supervise access to medical data. It is possible to implement mechanisms that enable controlled data deletion or restrict access to only authorised parties.
- The addition of an intermediary layer between the interface and the blockchain is also a viable alternative. This layer enables access control and administration of medical data, including deletion based on the fulfilment of certain conditions.
- Data anonymization: Instead of deleting data directly, identifying information can be removed using anonymization techniques. Thus, the data remains in the blockchain, but cannot be associated with particular individuals.

The first three techniques involve adding additional structures around blockchains. Storing the actual data off-chain raises the most concerns, as the purpose of putting them in a blockchain is to ensure their immutability. Placing them off-chain introduces the possibility of making changes and only complicates the structure. Smart contracts and special functions entail additional expenses, turning simple modifications into complex software solutions. The use of an intermediary layer has the same drawbacks—complexity, cost, and the potential loss of some advantages offered by the proposed solution. What seems to be the most reasonable approach for the hybrid model is data anonymization.

In this article, we examine a method for erasing patient data involving the encryption of identifying information and the ability to delete encryption keys. In addition to the aforementioned techniques, we will investigate the potential of blockchain tables that permit the eradication of particular information after an established amount of time and under specific conditions (subject to having the appropriate certificates).

Experiments

The database implementation phase utilised the Oracle server version 21c capabilities, including the blockchain table mechanism. The model has been implemented and populated with sample data. The anonymization process was tested through key deletion and direct deletion of partial data from the blockchain tables.

KEYWORDS: Healthcare hybrid database; Blockchains; Legal requirements for healthcare databases; Data Security.

REFERENCES

Azaria, A., Ekblaw, A., Vieira, T. , and Lippman, A. (2016). “Medrec: Using blockchain for medical data access and permission management,” in 2016 2nd international conference on open and big data (OBD). IEEE, 2016, pp. 25–30.

De Aguiar, E. J. , Faiçal, B. S., Krishnamachari, B. and Ueyama, J., (2020) “A survey of blockchain-based strategies for healthcare,” ACM Computing Surveys (CSUR), vol. 53, no. 2, pp. 1–27.

Farouk, A., Alahmadi, A., Ghose, S., and Mashatan, A., (2020) “Blockchain platform for industrial healthcare: Vision and future opportunities,” Computer Communications, vol. 154, pp. 223–235.

European Parliament and Council of the European Union. Consolidated text: 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 (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04

Linn, L. A., Koo, M. B. et al.,(2016) “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. Gaithersburg, Maryland, United States: ONC/NIST, pp. 1–10.

Rosen, J. (2011). The right to be forgotten. *Stan. L. Rev. Online*, *64*, 88.

Shahnaz, A., Qamar, U., and Khalid, A.,(2019) “Using blockchain for electronic health records,” IEEE access, vol. 7, pp. 147 782–147 795.

Yaqoob, I. ,Salah, K. Jayaraman, R. and Al-Hammadi, Y.,(2022) “Blockchain for healthcare data management: opportunities, challenges, and future recommendations,” Neural Computing and Applications, vol. 34, no. 14, pp. 11 475–11 490.

ENHANCING SECURITY GOVERNANCE IN MEDICAL DATABASES: A POLICY-BASED APPROACH WITH HYBRID RELATIONAL-BLOCKCHAIN MODEL

Olga Siedlecka-Lamch, Sabina Szymoniak

Department of Computer Science, Czestochowa University of Technology (Poland)

olga.siedlecka@icis.pcz.pl; sabina.szymoniak@icis.pcz.pl

EXTENDED ABSTRACT

Medical databases play a crucial role in managing patient information, clinical trials, medical histories, and other facets of healthcare in the modern era, due to the increasing number of patients and technological advancements. However, as access to enormous quantities of medical data increases, so do concerns about the security and privacy of this information.

In response to these challenges, this article proposes an innovative method for enhancing the security of medical databases by combining traditional relational databases with blockchain technology in a hybrid model. Our primary objective is to investigate how the hybrid model can enhance the security of medical databases by assuring data integrity, privacy protection, and access control.

This article will delve into the discussion of various security aspects of a hybrid-model-based medical database. We will introduce the concept of user certificates and the assignment of permissions, thereby facilitating data access management. In addition, we will investigate the various transaction security levels that safeguard data integrity and guarantee transaction immutability.

Implementing our solutions can contribute to creating secure medical databases and improving patient data security. The explicit organisation and immutability of data stored in blockchains can increase patient trust. Based on the hybrid model, the solutions analysed in this article can serve as a basis for future research and the implementation of innovative systems for managing medical data.

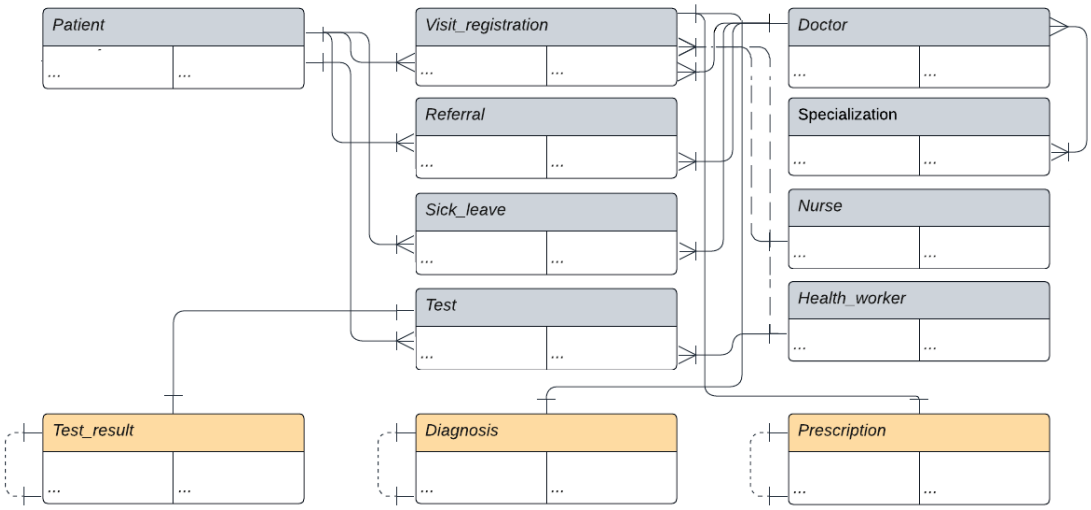
Related Work

Blockchain technology is widely used in many areas ((Tan et al., 2022), (Shi et al., 2022), (Khanna et al., 2022)). One of them is the healthcare industry. Information management, drug tracking, data security, and privacy were considered when using blockchain in the medical field. Farouk et al. reviewed the use of blockchain and IoT in healthcare systems (Farouk et al., 2020). The benefits of storing and exchanging health data via blockchain were demonstrated by Lin et al. (Lin et al., 2016). Yaqoob et al. demonstrated the viability of using blockchain in healthcare applications (Yaqoob et al., 2021). Jabbar et al.'s study concentrated on the difficulties and potential future paths of pharmaceutical supply chain intervention (Jabbar et al., 2021).

Database Model

In the examined model, we have incorporated patient details, visit history, medical leave records, test results, diagnoses, referrals, prescribed medications, medical staff information, disease codes, and their respective categories. Data directly associated with the treatment process (test results, diagnoses, prescribed medications) is stored in a blockchain (highlighted in orange in Figure 1). This results in an unalterable, observable, and easily analyzed sequence of events from every medical device. Patients also have access to this information; however, only doctors with the appropriate certificates assigned to their profiles can make modifications.

Figure 1. Simplified conceptual model

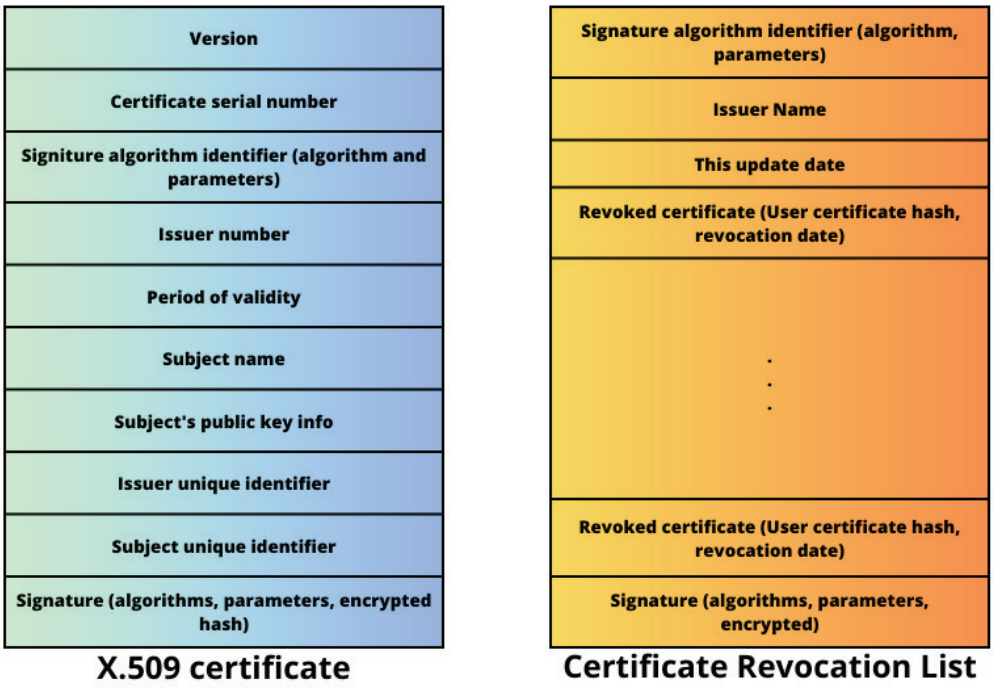


Certificates System

X.509 is a standard format for public key certificates widely used in various applications Cooper et al., 2008). These certificates combine cryptographic key pairs with identities that can be individuals, organizations, or websites. The organization can ensure its identity and exchange digitally signed messages thanks to them. The compromised X.509 certificates are immediately revoked by the Platform, which distributes a Certificate Revocation List in the network and prepares and sends new certificates. Also, Oracle Blockchain Platform used them to guarantee blockchain network security and data integrity.

Figure 2 shows the latest version of X.509 certificate and Certificate Revocation List structures. The certificate includes details about the certificate subject (subject's name, public key information, and unique identifier) and certificate issuer (issuer's number and unique identification). The certificate also includes details about the employed algorithm, signature, version, serial number, and validity period. The revoked certificates, the updated date, the signature, and the issuer's name are all listed in the certificate revocation list. Only users with issued certificates can insert data into the database's key tables.

Figure 2. X.509 certificate and Certificate Revocation List structures.



Security Aspects

Medical data security encompasses several key aspects within the security policy framework. The security policy is a formal document prepared for a specific entity, comprising analyses and directives pertaining to risk management and asset protection within that context. It entails the identification and analysis of various threats, such as terrorism, cyberattacks, criminal activities, or natural disasters. This analysis enables the establishment of priorities and the adaptation of appropriate protective measures. Objectives within the security domain will be defined based on these analyses, encompassing the safeguarding of life and property, ensuring individual stability, protecting information, and considering specific threats, strategies, and measures for achieving these goals. In addition, the security policy should define the organisational structure and responsibility for managing security, as well as define detailed procedures and guidelines for actions to be taken in the event of threats or incidents. These may include guidelines for responding to cyberattacks, disaster evacuation procedures, and guidelines for secure data storage. Furthermore, the security policy will establish a system for monitoring, evaluation, and continuous enhancement of security-related activities.

Furthermore, medical data security intersects with ethics due to its vulnerability. Security influences decision-making and actions concerning the safeguarding of data, resources, and the interests of the entity and individuals. Particularly in the case of medical data, the security policy needs to embody privacy and data confidentiality principles and operational procedures. It must detail how collected information is stored, processed, and made accessible, as well as the constraints and protocols related to this data. Additionally, the policy must guarantee adequate protection of personal data and ensure compliance with privacy regulations and standards.

From the technical point of view, we will consider such issues as system availability, authorization to the system, data access permissions, restoring regular operation after a failure

or incident and user permissions and services that can be used. In the system using medical data, we must consider three types of users' roles: administrators, medical staff and patients.

The administrators are responsible for managing the entire system, supervising access to data, configuring and maintenance of the infrastructure, including restoring the system after failures. We can specify server administrators, schema administrators, and blockchain and certificate administrators. The medical staff, like doctors or nurses, can access their patient's medical records (medical histories, test results, diagnoses, and prescribed medications). Their permissions are usually determined by speciality, area of practice, or other factors. Doctors and diagnosticians will have certificates to add data to blockchains. The patients use the system to obtain information about their health, medical history, test results, prescriptions and other data about their healthcare. Patients only have access to their data. They cannot edit their data.

Furthermore, when we use blockchain technology, also we must consider issues aimed at ensuring the protection of confidentiality, integrity and availability of information stored in the blockchain network. The specific security policy for blockchain data will be tailored to individual needs, but we can point out the following guidelines for this issue. The data stored on the blockchain is appropriately secured using robust encryption algorithms. The security policy defines how blockchain users are authenticated and authorized. The policy should define the rules for accessing data stored in the blockchain and includes regular updates of blockchain software and the use of security patches to address known vulnerabilities.

KEYWORDS: Database Security; Healthcare database design; Blockchains; Hybrid models.

REFERENCES

Cooper, D., Santesson, S., Farrell, S., Boeyen, S., Housley, R., & Polk, W. (2008). RFC 5280: Internet X. 509 public key infrastructure certificate and certificate revocation list (CRL) profile.

Farouk, A., Alahmadi, A., Ghose, S., & Mashatan, A. (2020). Blockchain platform for industrial healthcare: Vision and future opportunities. Computer Communications, 154, 223-235.

Khanna, A., Sah, A., Bolshev, V., Burgio, A., Panchenko, V., & Jasiński, M. (2022). Blockchain–Cloud Integration: A Survey. Sensors, 22(14), 5238.

Linn, L. A., & Koo, M. B. (2016, September). 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. Gaithersburg, Maryland, United States: ONC/NIST (pp. 1-10).

Shi, Z., Zhou, H., de Laat, C., & Zhao, Z. (2022). A bayesian game-enhanced auction model for federated cloud services using blockchain. Future Generation Computer Systems, 136, 49-66.

Tan, W., Zhu, H., Tan, J., Zhao, Y., Xu, L. D., & Guo, K. (2022). A novel service level agreement model using blockchain and smart contract for cloud manufacturing in industry 4.0. Enterprise Information Systems, 16(12), 1939426.

Yaqoob, I., Salah, K., Jayaraman, R., & Al-Hammadi, Y. (2021). Blockchain for healthcare data management: opportunities, challenges, and future recommendations. Neural Computing and Applications, 1-16.

ETHICAL THREATS ASSOCIATED WITH THE APPLICATION OF ARTIFICIAL INTELLIGENCE: A COMPREHENSIVE REVIEW

Sabina Szymoniak, Mariusz Kubanek

Department of Computer Science, Czestochowa University of Technology, Poland

mariusz.kubanek@icis.pcz.pl; sabina.szymoniak@icis.pcz.pl

EXTENDED ABSTRACT

Artificial Intelligence (AI) has witnessed unprecedented growth in recent years, revolutionizing various industries and domains. However, along with its immense potential, the widespread adoption of AI also raises profound ethical concerns. This comprehensive scientific review article aims to explore the ethical threats associated with the application of AI, focusing on recent and relevant articles published in 2020 and onwards. By incorporating these citations, this review provides an in-depth analysis of the emerging ethical challenges that necessitate careful consideration and proactive measures.

The review begins by examining the issue of bias and discrimination in AI systems. Research by Angwin et al. (2016) reveals the presence of bias in predictive models used within criminal justice systems, leading to disparate outcomes for different racial and ethnic groups. The study highlights the alarming implications of such biases, as they perpetuate systemic inequalities and hinder the fairness of the criminal justice system. To address this concern, recent studies have emphasized the importance of developing fair and unbiased algorithms through careful data selection and algorithmic design. Mittelstadt et al. (2016) delve into the ethics of algorithms and the ongoing debate surrounding their fairness and accountability. They emphasize the need for transparent decision-making processes and comprehensive audits to detect and rectify biases in AI systems. Furthermore, a study by Wachter, Mittelstadt, and Floridi (2020) focuses on the importance of transparency, explainability, and accountability in AI systems for robotics. They argue that the development of AI algorithms should incorporate transparency mechanisms that enable users to understand the decision-making process. This not only ensures fairness but also fosters trust between users and AI systems. In addition to algorithmic fairness, ethical guidelines and regulations play a crucial role in addressing bias and discrimination. Jobin, Ienca, and Vayena (2020) discuss the global landscape of ethics guidelines in biomedicine and highlight the need for comprehensive AI ethics frameworks. These frameworks provide guidance on ensuring fairness, non-discrimination, and inclusivity in AI applications across various domains. By integrating these findings, it becomes evident that combatting bias and discrimination in AI systems requires a multi-faceted approach. This includes data-driven approaches, algorithmic transparency, and the implementation of robust ethical frameworks to govern the development and deployment of AI technologies.

Another critical area of ethical concern is privacy and data protection in AI applications. Jobin, Ienca, and Vayena (2019) shed light on the global landscape of AI ethics guidelines, highlighting the need for robust privacy-preserving techniques and data anonymization methods to safeguard individuals' privacy rights. They emphasize the significance of adopting privacy-centric approaches in AI development to address the potential risks associated with the collection, storage, and processing of personal data. In line with these concerns, ethical frameworks

proposed by Floridi et al. (2018) emphasize the importance of incorporating privacy-enhancing measures into AI systems. They argue that privacy should be treated as a foundational value throughout the entire life cycle of AI technologies. The authors suggest adopting privacy by design principles, data minimization strategies, and the implementation of strong encryption techniques to mitigate privacy risks. Moreover, research by Dignum and Marchiori (2021) explores the ethical challenges surrounding privacy in AI systems. They discuss the tensions between privacy and AI, highlighting the potential trade-offs that arise when leveraging personal data for AI-driven applications. The study emphasizes the need for clear regulations and guidelines to strike a balance between the benefits of AI and the protection of individuals' privacy rights. Additionally, Bostrom (2014) addresses the ethical implications of data protection in the context of superintelligence. He underscores the significance of safeguarding sensitive information and preventing unauthorized access, as superintelligent AI systems could pose unprecedented risks if they were to gain access to vast amounts of personal data. These studies collectively underscore the importance of privacy and data protection in the ethical deployment of AI systems. By adopting privacy-enhancing techniques, data anonymization methods, and incorporating privacy as a foundational principle, we can work towards ensuring the responsible and ethical use of AI technologies while respecting individuals' privacy rights.

Accountability and transparency in AI decision-making processes are vital to maintain public trust and ensure responsible AI deployment. Calo (2017) stresses the need for ethical guidelines and regulations that promote transparency, explainability, and accountability in AI algorithms and systems. The author emphasizes that transparency is essential to enable users and stakeholders to understand how AI systems arrive at their decisions and to identify potential biases or errors. Furthermore, Bostrom (2014) discusses the challenges associated with ensuring the accountability of AI systems and proposes strategies to mitigate risks. He argues that as AI systems become more autonomous and capable of making decisions with far-reaching consequences, it is crucial to establish mechanisms for holding these systems accountable for their actions. Bostrom suggests the development of certification and auditing processes, as well as the creation of regulatory frameworks, to ensure that AI systems are designed, developed, and deployed in an accountable manner. In addition to the aforementioned works, research by Mittelstadt et al. (2020) delves into the importance of transparent and explainable AI for robotics. The authors emphasize that AI systems should provide clear explanations for their decisions to enhance accountability and enable human users to assess their reliability. They propose approaches such as interpretable machine learning and algorithmic explanations to address the challenges of accountability and transparency in AI systems. The incorporation of accountability and transparency measures is essential not only for ethical considerations but also to address potential societal, legal, and regulatory challenges posed by AI technologies. By ensuring that AI systems are accountable for their actions and that their decision-making processes are transparent and explainable, we can foster public trust and confidence in the responsible deployment of AI.

The societal impact of AI on the workforce is another significant ethical concern. O'Neil (2016) explores how AI and automation technologies can lead to job displacement and widen socioeconomic inequalities. The author highlights the potential consequences of these technologies, particularly in sectors where AI systems can perform tasks more efficiently and cost-effectively than human workers. The displacement of workers in these sectors can result in unemployment, income disparities, and social unrest. Mitigating these impacts requires proactive measures to address the ethical and social implications of AI-driven automation.

Reskilling and upskilling programs play a crucial role in preparing the workforce for the changing job landscape. By providing individuals with the necessary skills and knowledge to adapt to emerging technologies, these programs can enable workers to transition into new roles and industries that are less susceptible to automation. Policy interventions also play a vital role in ensuring a just transition for affected individuals. Governments and regulatory bodies need to develop policies and initiatives that address the socioeconomic consequences of AI-driven automation. This may include measures such as income support, retraining programs, and job creation efforts in emerging industries. Furthermore, research by Russell and Norvig (2016) discusses the potential long-term impacts of AI on the workforce and emphasizes the importance of considering societal implications when designing AI systems. The authors argue for a responsible and human-centric approach to AI development that takes into account the broader social and economic context. By recognizing the potential impacts on the workforce and implementing measures to address these concerns, we can strive for a future where AI technologies contribute to socioeconomic progress without leaving behind vulnerable individuals or exacerbating existing inequalities.

Moreover, the development and deployment of autonomous systems, including autonomous weapon systems, pose critical ethical dilemmas. Russell and Norvig (2016) highlight the need for clear ethical guidelines and regulations to govern the use of AI in military applications and ensure human oversight in critical decision-making processes. They emphasize the importance of maintaining human control over autonomous systems to prevent the escalation of conflicts and minimize the risks associated with the uncontrolled use of AI technologies in warfare. The ethical concerns surrounding autonomous weapon systems have also been addressed by Arkin (2019). He argues for the development of ethical governor architectures that can ensure compliance with international humanitarian laws and ethical principles in the use of AI technologies in military contexts. The author stresses the necessity of integrating ethical considerations into the design and deployment of autonomous weapon systems to prevent unintended harm and promote responsible use. Additionally, research by Wang, Zhang, and Zhang (2020) focuses on the ethical challenges associated with the deployment of AI in military operations. The authors discuss the implications of AI-driven decision-making processes in warfare and highlight the importance of maintaining human accountability and responsibility for the actions of autonomous systems. They propose the integration of human-in-the-loop mechanisms to ensure that critical decisions are made by human operators rather than solely relying on AI algorithms. The development of clear ethical guidelines and regulations for the use of AI in military applications is crucial to address the ethical dilemmas posed by autonomous systems. By emphasizing human oversight, compliance with international laws, and the integration of ethical considerations into the design and deployment of AI technologies in the military, we can strive for the responsible and ethical use of AI in warfare.

In conclusion, the ethical considerations surrounding the application of AI demand careful attention. By reviewing recent articles published in 2020 and onwards, this comprehensive scientific review highlights the emerging ethical threats associated with AI. The findings underscore the importance of addressing issues related to bias and discrimination, privacy and data protection, accountability and transparency, workforce implications, and autonomous systems. Policymakers, researchers, and practitioners must collaborate to develop ethical frameworks and guidelines that guide the responsible and ethical development, deployment, and governance of AI.

KEYWORDS: Artificial Intelligence, ethics, data protection, autonomous systems.

REFERENCES

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

Wachter, S., Mittelstadt, B. D., & Floridi, L. (2020). Transparent, explainable, and accountable AI for robotics. Science Robotics, 5(47), eaaz8238. https://doi.org/10.1126/scirobotics.aaz8238

Jobin, A., Ienca, M., & Vayena, E. (2020). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 2(9), 389-399. https://doi.org/10.1038/s42256-020-0214-1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Prins, C. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Dignum, V., & Marchiori, E. (2021). Privacy and AI: Tensions and potential trade-offs. Minds and Machines, 31(1), 57-71. https://doi.org/10.1007/s11023-020-09541-7

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Calo, R. (2017). Artificial intelligence policy: A primer and roadmap. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2972850

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

Russell, S., & Norvig, P. (2016). Artificial intelligence: A modern approach (3rd ed.). Pearson.

Arkin, R. C. (2019). Ethical governance of lethal autonomous systems. In The Oxford handbook of ethics of AI (pp. 537-556). Oxford University Press.

Wang, Y., Zhang, H., & Zhang, J. (2020). The ethical challenges of using AI in military operations. AI & Society, 35(3), 735-744. https://doi.org/10.1007/s00146-019-00883-y

Alsaeed, N. H., & Nadeem, F. (2022). Authentication in the Internet of Medical Things: Taxonomy, Review, and Open Issues. *Applied Sciences*, *12*(15), 7487. https://doi.org/10.3390/app12157487

Khan, F., Xu, Z., Sun, J., Khan, F. H., Ahmed, A., & Zhao, Y. (2022). Recent Advances in Sensors for Fire Detection. *Sensors*, *22*(9), 3310. https://doi.org/10.3390/s22093310

Masud, M., Gaba, G. S., Kumar, P., & Gurtov, A. (2022). A user-centric privacy-preserving authentication protocol for IoT-AmI environments. *Computer Communications*, *196*, 45–54. https://doi.org/10.1016/j.comcom.2022.09.021

ETHICS IN INTERNET OF THINGS: CHALLENGES AND OPPORTUNITIES

Sabina Szymoniak, Mariusz Kubanek

Department of Computer Science, Czestochowa University of Technology, Poland

sabina.szymoniak@icis.pcz.pl; mariusz.kubanek@icis.pcz.pl

EXTENDED ABSTRACT

The Internet of Things (IoT) is a network of connected physical devices. Devices exchange data between them using the Internet. IoT is the concept that connects different devices like home appliances, vehicles, sensors or smartphones to the internet network. IoT devices and connections exist in many areas (Szymoniak & Kesar, 2022). We utilize smart washing machines, TVs, and light bulbs. Thus, we can discover IoT gadgets in our daily lives. These gadgets use the proper sensors to regulate a building's lighting or water heating intelligently. With the aid of tracking gadgets, they can also safeguard our security (Khan et al., 2022; Alsaeed & Nadeem, 2022). Devices used in medical IoT assist in managing the critical functions of patients with chronic illnesses, testing blood glucose levels in people with diabetes, alerting doctors when a patient needs medication, and promptly delivering it to the patient (Singh et al., 2022). One of the common uses for IoT in the sector is to warn people about the potential for an earthquake (Sivakumar et al., 2022). In order to avoid potentially fatal scenarios, athletes might use IoT to regulate vital processes and performance (Zhou et al., 2021).

As mentioned, IoT devices use the Internet to communicate. Basically, they use wireless data transmission, for example, WiFi, and LTE / 5G, as secure channels supported by secure cryptographic protocols like SSL/TLS. However, IoT connections also implement and realize other security protocols specially designed for these solutions in the specific solutions. The security protocols define the order in which messages must be sent. We can indicate many security protocols dedicated to different solutions, for example, in medicine or healthcare (Rasslan et al., 2022), (Masud et al., 2022), in fog or edge processing (Pardeshi et al., 2022), for industry (Yi et al., 2022), for meetings (Szymoniak & Siedlecka-Lamch, 2022) or suitable for many domains (Yan et al., 2022).

Depending on the protocol's application, we send many different data during communication between devices. Each security protocol should implement the so-called CIA triad, the basic IT security concept. CIA triad ensures the protection of information. Achieving a balance between its three goals is crucial to effectively securing systems and data. CIA triad goals are confidentiality, integrity and availability. Confidentiality ensures that information will be available only to authorized users and protects against unauthorized access. The integrity ensures that data will be accurate, unaltered and undamaged and prevent data modifications or deletions by unauthorized users. The availability ensures that information is available for users at the requested time when they want it. Using backup resources and appropriate hardware safeguards improves availability (Szymoniak & Kesar, 2022).

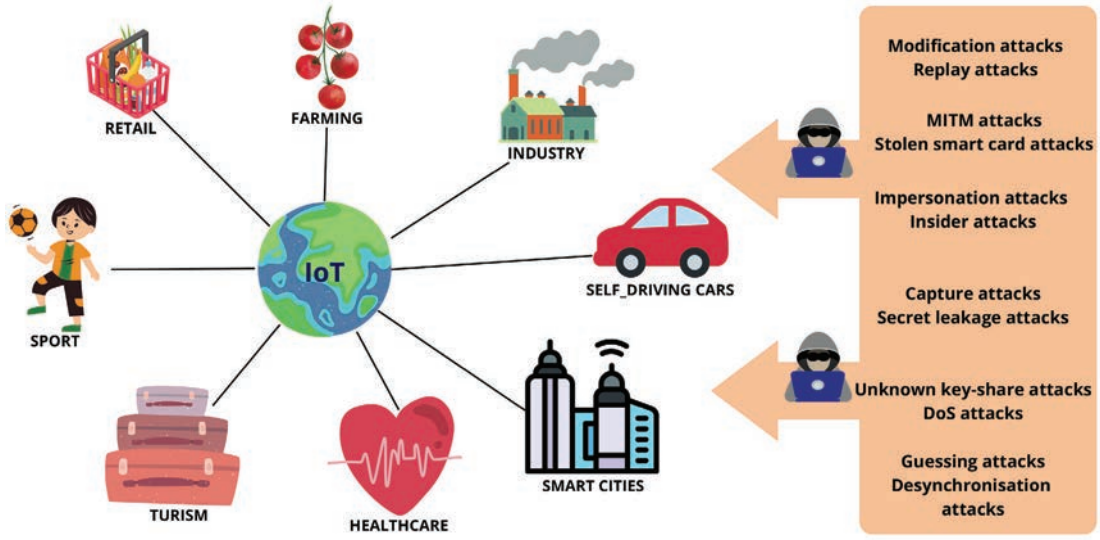
Also, the security protocols should satisfy some security features. The first is mutual authentication, which refers to two users verifying their authority over each other. User anonymity provides that the user’s authority will be anonymous or hidden. Next, the perfect forward secrecy ensures that even if a private key is compromised in the future, previously

mentioned secret keys will not be exposed or compromised. The perfect backward secrecy ensures that even the private keys have been compromised in the past and does not allow previous sessions to be compromised. The last security feature is untraceability. This feature ensures that activities or transactions cannot be traced back to a specific user ((Szymoniak & Kesar, 2022), (Kubanek et al., 2022)).

Unfortunately, the security protocols, even if they fulfil these features, can be vulnerable to many attacks by malicious users ((Szymoniak et al., 2017), (Szymoniak et al., 2018)). From many statistics, there are more than 2000 cyberattacks per day. The attackers search for vulnerabilities in such systems and try to break into them. The system hacking effects are hazardous for many reasons. First, the users can lose their devices because the attacker obtains control of them. Next, he can try to eavesdrop on whole communication in the network and steal private data, logins or passwords. Moreover, the attacker can take control of other devices in the network or the whole Smart Home. Figure 1 summarises IoT solutions and typical cyberattacks on IoT systems.

The cyberattacks' influence on users and their data upon IoT systems entails the ethical consideration of communication on such systems. We must think about the ethics of data storage, which answers questions like what data can be stored, what data should not be stored, and what is the maximum time necessary for sensitive data storage. It is necessary because stored data can be stolen from devices or servers and used. Also, we must consider the risk of data leakage from IoT systems. Moreover, in the case of security protocols, we must investigate how they deal with mentioned security features, what communication elements make vulnerabilities, and how to protect IoT systems and their users against cyberattacks. This paper will consider the challenges and opportunities of ethics in the Internet of Things systems.

Figure 1. IoT solutions and typical cyberattacks on IoT systems.



KEYWORDS: Internet of Things, ethics, security, attacks, vulnerabilities.

REFERENCES

Alsaeed, N. H., & Nadeem, F. (2022). Authentication in the Internet of Medical Things: Taxonomy, Review, and Open Issues. *Applied Sciences*, *12*(15), 7487. https://doi.org/10.3390/app12157487

Khan, F., Xu, Z., Sun, J., Khan, F. H., Ahmed, A., & Zhao, Y. (2022). Recent Advances in Sensors for Fire Detection. *Sensors*, *22*(9), 3310. https://doi.org/10.3390/s22093310

Kubanek, M., Bobulski, J., & Karbowiak, Ł. (2022). Intelligent Identity Authentication, Using Face and Behavior Analysis. *ETHICOMP 2022*, 42.

Masud, M., Gaba, G. S., Kumar, P., & Gurtov, A. (2022). A user-centric privacy-preserving authentication protocol for IoT-AmI environments. *Computer Communications*, *196*, 45-54. https://doi.org/10.1016/j.comcom.2022.09.021

Pardeshi, M. S., Sheu, R., & Yuan, S. (2022). Hash-Chain Fog/Edge: A Mode-Based Hash-Chain for Secured Mutual Authentication Protocol Using Zero-Knowledge Proofs in Fog/Edge. *Sensors*, *22*(2), 607. https://doi.org/10.3390/s22020607

Rasslan, M., Nasreldin, M., & Aslan, H. K. (2022). Ibn Sina: A patient privacy-preserving authentication protocol in medical internet of things. *Computers & Security*, *119*, 102753. https://doi.org/10.1016/j.cose.2022.102753

Singh, S., Nandan, A. S., Sikka, G., Malik, A., & Vidyarthi, A. (2022). A secure energy-efficient routing protocol for disease data transmission using IoMT. *Computers & Electrical Engineering*, *101*, 108113. https://doi.org/10.1016/j.compeleceng.2022.108113

Sivakumar, P., Sandhya Devi, R.S., Ashwin, M., Rajan Singaravel, M.M. & Buvanesswaran, A.D. (2022). Protocol Design for Earthquake Alert and Evacuation in Smart Buildings. In: Rani, S., Sai, V., Maheswar, R. (eds) IoT and WSN based Smart Cities: A Machine Learning Perspective. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-84182-9_1

Szymoniak, S., & Kesar, S. (2022). Key Agreement and Authentication Protocols in the Internet of Things: A Survey. *Applied Sciences*, *13*(1), 404. https://doi.org/10.3390/app13010404

Szymoniak, S., & Siedlecka-Lamch, O. (2022). Securing Meetings in D2D IoT Systems. *ETHICOMP 2022*, 31.

Szymoniak, S., Siedlecka-Lamch, O., & Kurkowski, M. (2017). Timed analysis of security protocols. In *Information Systems Architecture and Technology: Proceedings of 37th International Conference on Information Systems Architecture and Technology–ISAT 2016–Part II* (pp. 53-63). Springer International Publishing.

Szymoniak, S., Siedlecka-Lamch, O., & Kurkowski, M. (2018). On some time aspects in security protocols analysis. In *Computer Networks: 25th International Conference, CN 2018, Gliwice, Poland, June 19-22, 2018, Proceedings 25* (pp. 344-356). Springer International Publishing.

Yan, D., Luo, Y., Chen, X., Tong, F., Xu, Y., Tao, J., & Cheng, G. (2022). A Lightweight Authentication Scheme Based on Consortium Blockchain for Cross-Domain IoT. *Security and Communication Networks*, *2022*, 1-15. https://doi.org/10.1155/2022/9686049

Yi, F., Zhang, L., Xu, L., Yang, S., Lu, Y., & Zhao, D. (2022). WSNEAP: An Efficient Authentication Protocol for IIoT-Oriented Wireless Sensor Networks. *Sensors*, *22*(19), 7413. https://doi.org/10.3390/s22197413

Zhou, H., Wang, Z., Zhao, W., Tong, X., Jin, X., Zhang, X., Yu, Y., Liu, H., Ma, Y., Li, S., & Chen, W. (2021). Robust and sensitive pressure/strain sensors from solution processable composite hydrogels enhanced by hollow-structured conducting polymers. *Chemical Engineering Journal*, *403*, 126307. https://doi.org/10.1016/j.cej.2020.126307

THEORETICAL FRAMEWORK USING AI: IMPROVING SERVICES WITHIN SMART CITIES

Sabina Szymoniak, Shalini Kesar

Czestochowa University of Technology (Poland), Southern Utah University (USA)

sabina.szymoniak@icis.pcz.pl; kesar@suu.edu

EXTENDED ABSTRACT

This paper is part of an on-going collaborative research to develop a framework that will support notifying emergency services within smart cities ((Joshi et al., 2016), (Tura et al., 2022)). The framework is designed to provide a support system for existing emergency services like ambulances within the city. After reviewing the challenges of the existing frameworks linked with artificial intelligence, the authors propose a theoretical framework using AI that overcomes the existing challenges to provide an efficient mechanism for ambulance services within smart cities. The collaborative work of the authors, experts in risk management and security of computer systems, will provide a significant contribution in the research area that combines best practices of cybersecurity and smart cities. Given that smart cities are increasingly becoming popular in urban areas, this framework, an on-going research, can be a starting point for many services that can help in mitigating, minimising, managing as well as transferring risks when it comes to human life.

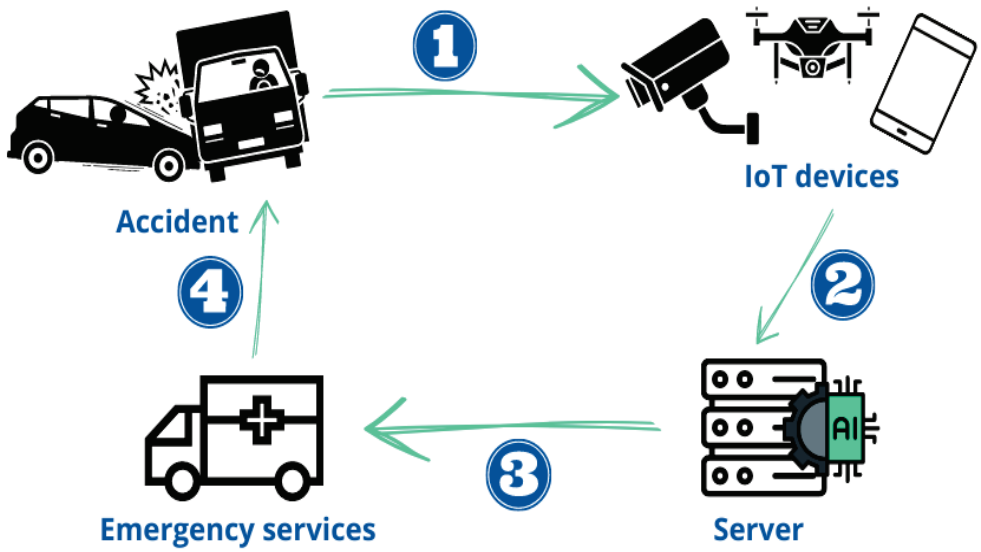
Our daily lives cannot function without smart devices. We employ a variety of gadgets, like intelligent refrigerators, vacuum cleaners, and ovens, to carry out preprogrammed tasks automatically and share data. These gadgets are controlled by smartphones, various sensors, and software that enables us to manage a working environment and carry out particular tasks without human participation. Using such gadgets, we can control our home from anywhere globally, maintaining the right room temperature and ensuring their security. Such devices belong to the Internet of Things, IoT for short. Moreover, they can be used for many more advanced tasks connected with human safety, especially when they are equipped with Artificial Intelligence (AI) methods (("Internet of Things," 2022), (Szymoniak & Kesar, 2022)).

Dependency on data and technology in smart cities will continue to increase. A recent article in IT Magazine (2023) states that data by smart cities is expected to grow by more than 140% between 2023 and 2027. More so, there will be more cellular connections in the Internet of Things projects in smart cities, which are expected to increase at a compound annual rate of 17.9% between 2022 and 2027, reaching a plateau of more than 122 million, with particularly high growth in the next two years. As a result, there will also be more risks from data breaches to fatal accidents on the road to minimise, manage, mitigate as well transfer risks, rescue services dependency will increase to help the injured and secure the area around the incident. Also, city dwellers can witness situations that may turn into dangerous situations, for example, when a group of people argue. In some cases, the argument may turn into a fight.

This paper proposes a proof of concept as a framework (see Figure 1 below) that focuses on using a network of IoT devices as an intelligent system to support ambulance services, which can minimise fatality. As shown below, smart devices equipped with a camera can capture the

moment of an accident or other dangerous situation and then send photos to a trusted server equipped with AI-based software to recognise the situation type, decide if the situation is dangerous, and notify the rescue services.

Figure 1. The architecture of the proposed system.



As mentioned earlier, Figure 1 shows the proposed system's architecture, which consists of four ingredients. The first is the scene of a dangerous situation, like an accident. The second is the network of IoT devices. The third is the trusted server equipped with the appropriate software. The last ingredient is the rescue services. The operation of the system will be a continuous, four-step process. The situation happens in the first step, and the IoT device takes this event's photo. Next, the device sends the captured photo to the trusted server via the Internet (step 2). After that, the server will process the obtained photos using AI-based software and decide whether the situation is dangerous. If the reported situation is dangerous, the server will notify the rescue services immediately (step 3). In the fourth step, rescue services will help injured victims or secure the area.

Many types of IoT devices equipped with cameras can be used for this system. Also, we can employ users and their smartphones in it. The whole process of system operation should satisfy some security requirements. It should implement and realise the appropriate security protocol for communication between devices connected to the system (including the trusted server and the rescue services). The security protocol should guarantee high security in inter-entity communication, including scalability, authenticity, assault resistance, and data confidentiality. Ensuring that unauthorised parties cannot access the sent information is connected to data confidentiality. Ensuring data is not altered or lost while in transit entails maintaining data integrity. Verifying the identification of users or communication systems is referred to as authenticity. Attack resistance protects users and their data from various network threats. Scalability is the ability to securely communicate with many users or systems while accommodating the addition of new users or systems without requiring a complete protocol change. The protocol should also incorporate AAA (Authentication, Authorization, Accounting)

logic, whose elements govern user identification within the network, enforce user rules, and log session statistics (Steingartner et al., 2022).

Such a system involves risks, for example, associated with security, privacy, data storage, or AI use. Security and privacy risks are associated with many threats from computer networks. Each computer system is the target of cyberattacks. Hackers have many abilities and tools to break into the computer system, steal users' data and then use them in an unethical way. So the users can lose their privacy.

The risk of storing data on servers is essential to using artificial intelligence and information technologies. If the data stored on the servers contain personal information, there is a risk of unauthorised access or use by third parties. Cyberattacks, data leaks or inadequate security measures can violate users' privacy. Servers that store data are at risk of mentioned cyberattacks. Hackers may try to take control of servers or steal stored data for illegal use, such as identity theft or blackmail. Regardless of the cause (hardware failure, human error, attacks), there is a risk of losing server data. If proper backup and data redundancy strategies are not in place, a server failure can permanently lose valuable information. Storing data on servers requires proper management. Configuration errors, insufficient security measures or improper procedures can lead to unauthorised access, loss or accidental disclosure of data. Storing data on servers requires compliance with relevant laws and regulations, such as the General Data Protection Regulation (Voigt & Von Dem Bussche, 2017) in the European Union. Failure to comply with these requirements may lead to legal consequences, financial penalties and loss of user trust.

Using artificial intelligence (AI) in such systems carries certain risks. First, AI can make mistakes or produce unpredictable results. Learning algorithms may base their decisions on training data that may be incomplete, error-prone, or biased. This can lead to incorrect or unfair decisions. If systems are wholly dependent on AI, failures, programming errors, or technical issues can cause severe disruptions in the functioning of these systems. This can have negative consequences for society. AI can pose ethical and responsibility challenges. Decisions made by AI systems can have profound social impacts. We must be sure that AI's decision about dangerous situations is correct and will not cause human death.

To conclude, this on-going research is highlighted in this paper, where it discusses the architecture of the proposed system and its requirements, challenges and risk. This framework has considered many factors, such as previous research outcomes of the author, existing frameworks in this context, most importantly, the need and requirements for a safe, functional smart city. Given that there is lack or no such proposed framework, this is a significant contribution that can be used as a starting framework for other contexts.

KEYWORDS: Smart cities, AI, ambulance services, risk and security.

REFERENCES

Internet of Things. (2022). In Transactions on Computer Systems and Networks. Springer Nature. https://doi.org/10.1007/978-981-19-1585-7

Joshi, S., Saxena, S., & Godbole, T. (2016). Developing smart cities: An integrated framework. *Procedia Computer Science*, *93*, 902-909.

Steingartner, W., Možnik, D., & Galinec, D. (2022, November). Disinformation Campaigns and Resilience in Hybrid Threats Conceptual Model. In *2022 IEEE 16th International Scientific Conference on Informatics (Informatics)* (pp. 287-292). IEEE.

Szymoniak, S., & Kesar, S. (2022). Key Agreement and Authentication Protocols in the Internet of Things: A Survey. Applied Sciences, 13(1), 404. https://doi.org/10.3390/app13010404

Tura, N., & Ojanen, V. (2022). Sustainability-oriented innovations in smart cities: A systematic review and emerging themes. *Cities*, 103716.

Voigt, P., & Von Dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR). https://doi.org/10.1007/978-3-319-57959-7

NATIONAL CYBERSECURITY STRATEGY ACTION PLAN FOR CYBER RESILIENCE: QUALITATIVE DATA AND ACHIEVEMENTS

William Steingartner, Darko Galinec

Technical University of Košice (Slovakia), Zagreb University of Applied Sciences (Croatia)

william.steingartner@tuke.sk; darko.galinec@tvz.hr

EXTENDED ABSTRACT

Cyber issues of importance to the state and the global environment represent a much wider area than the field of cybersecurity and are closely related to several traditional departments of public administration. Cybersecurity in these matters is the basis for their smooth development in the virtual dimension of modern society. Cybersecurity is a part of all public administration processes, as all processes rely on the proper functioning of communication and information systems, either directly, through data processing, storage, and transmission, or in directly through the management of basic services (e.g., electricity distribution, transport, etc.). Given the widespread dispersion of responsibilities of state bodies in cyberspace, the establishment of the National Council for Cybersecurity, Operational and Technical Coordination for Cybersecurity and the development of the National Cyber Security Strategy and Action Plan for its implementation establishes a mechanism for sharing information and harmonizing public administration professional and political/administrative level. This paper presents a qualitative assessment of the implementation of the Action Plan of the Strategy based on the outcomes of reporting to the holders and co-carriers of the implementation of the Action Plan's measures at the state level.

In the development of the National Cybersecurity Strategy and Action Plan for its implementation comprehensive approach to cybersecurity by covering cyberspace and infrastructure and users that fall under the jurisdiction of the Republic of Croatia (citizenship, registration, domain, address) is used as well as integration and harmonization of activities and measures arising from various aspects of cybersecurity and falling under the competence of various organizations and their complementarity in order to create a safer common cyberspace.

A proactive approach by constantly adapting the activities and measures applied in cyberspace and by occasionally adapting the relevant strategic frameworks was needed for strengthening the resilience, reliability, and adaptability of information systems by implementing certification, accreditation, and security protocols (Szymoniak, 2021a; Szymoniak, 2021b), especially taking into account the specific requirements of data, services and other business processes on information systems. Using probabilistic techniques, various parameters and behaviors of security protocols embedded in the authentication systems can be thoroughly examined (Siedlecka-Lamch, 2020). The basic principles on which modern society is based (Cesarec, 2020; Gálik & Tolnaiová, 2019) are also applied in the cyberspace that makes up the virtual dimension of society:

- Application of the law for the purpose of protection of human rights and freedoms, especially privacy and the right to expression, property, and all other essential features of an organized modern society.

- Harmonized legislative framework and continuous improvement of regulatory mechanisms through harmonized initiatives of all sectors of society, i.e., bodies and legal entities.
- The principle of subsidiarity through the systematic elaboration of the power to decide and inform on cybersecurity issues to the body whose competence largely covers the problem to be solved, whether the problem relates to the organization, coordination and cooperation, or technical capabilities to respond to computer communication threats and information infrastructure.
- The principle of proportionality between the increase of protection measures and responsibilities and decreasing negative consequences (Tokarčíková et al., 2014) and accompanying costs and reduction of associated risks, i.e., greater possibilities to limit the threats that cause them.

Adopting a national cybersecurity strategy is one of the most important first steps in securing the national cyber infrastructure and services upon which the digital future and economic wellbeing of a modern nation depend (Spidalieri, 2017). Due to the ever-increasing availability and variety of sophisticated malicious digital tools and the ease with which these tools can be deployed, cybersecurity is now a crucial element of national security. Within this larger context, the concept of cyber defense, with its implicit military connotation, has also gained significantly more prominence (Dewar, 2018). In the following parts, we focus on countries which, in view of our best knowledge, we have found to have clear explanations key policy principles on cybersecurity, cyber defense and cyber resilience as essential concepts.

Cybersecurity and cyber defense are constantly shifting and evolving topics. The technology used to carry out cyber-attacks, and the tools required to mitigate or deter those attacks, is in a constant state of development and innovation. As a result, national policy relating to these topics also undergoes periodic shifts and changes, depending on national priorities (Dewar, 2018).

The National Cybersecurity Strategy planning and the Action Plan for its implementation development have been created and executed based on integration, inclusiveness and integrity principles by drafting strategic guidelines, concept development, plan development and plan assessment to achieve situational awareness at the national level. Furthermore, one of the main principles of the strategy was strengthening resilience, reliability and adjustability by applying universal criteria of confidentiality, integrity and availability of certain groups of information and recognized social values, in addition to complying with the appropriate obligations related to the protection of privacy, as well as confidentiality, integrity and availability for certain groups of information, including the implementation of appropriate certification and accreditation of different kinds of devices and systems, and also business processes in which such information is used (Dewar, 2018).

According to Gartner in (Top Priorities, 2020), security and risk management leaders are key enablers of digital business and are accountable for helping the enterprise balance the associated risks and benefits. By 2023, 30% of chief information security officers' effectiveness will be directly measured on the role's ability to create value for the business.

Three trends are making the highest impact for security and risk management leaders to be effective in their role and deliver business value to their organizations (Top Priorities, 2020):

- Citizen computing accelerates. Citizen computing is when a user creates new business applications using development and run time environments approved by IT. However, it's generally outside of IT visibility and traditional enforcement, which creates complexities for security and risk leaders tasked with protecting the organization.
- New digital initiatives create challenges. The security team is often not consulted until digital plans for the organization are well underway. In addition to reorienting the security program to address new technologies, effective security leaders are working with the board and business leaders to manage cyber-risk control expectations.
- Cybersecurity mesh emerges as the preferred delivery model for security services. This cloud-based and highly modular architecture makes it much more practical to control the uncontrollable. Cybersecurity mesh is the most efficient and effective way to extend security policy to digital assets that are outside of the traditional enterprise.

The main goal of this paper is to present the results of the implementation of the National Cyber Security Strategy because of research by qualitative analysis based on reports of sectoral bodies as responsible bodies for the implementation of action plan measures and implementation of the strategy.

KEYWORDS: Action plan, cyber attack, cyber defense, cyber resilience, national cybersecurity strategy, qualitative assessment.

REFERENCES

Arias-Oliva, M., Pelegrín-Borondo, J., & Matías-Clavero, G. (2019). Variables Influencing Cryptocurrency Use: A Technology Acceptance Model in Spain. *Frontiers in Psychology*, 10. https://doi.org/10.3389/fpsyg.2019.00475

Cesarec, I. (2020). Beyond physical threats: Cyber-attacks on critical infrastructure as a challenge of changing security environment – overview of cyber-security legislation and implementation in SEE countries. *Annals of Disaster Risk Sciences*, 3(1), 2020.

Dewar, R.S. (2018). National Cybersecurity and Cyberdefense Policy Snapshots: Collection 1. Center for Security Studies (CSS), ETH Zürich.

Gálik, S., & Tolnaiová, S.G. (2019). Cyberspace as a New Existential Dimension of Man, Cyberspace, Eds. E. Abu-Taieh, A. E. Mouatasim, and I. H. A. Hadid. IntechOpen, Rijeka, 2019, chap 2.

Siedlecka-Lamch, O. (2020). Probabilistic and Timed Analysis of Security Protocols. 13th International Conference on Computational Intelligence in Security for Informa tion Systems (CISIS 2020). Advances in Intelligent Systems and Computing, vol 1267, Eds. Herrero, Á., Cambra, C., Urda, D., Sedano, J., Quintián, H., & Corchado, E. Springer, 2021, p. 142–151.

Spidalieri, F. (2017). Italy: Building a Cyber Resilient Society, available online (accessed on 2022-04-19). Retrieved from https://www.ispionline.it/it/pubblicazione/italy-building-cyber-resilient-society-18229

Szymoniak, S. (2021a). Using A Security Protocol To Protect Against False Links, Moving technology ethics at the forefront of society, organisations and governments ETHICOMP

Book Series, Eds. Jorge Pelegrín Borondo, Mario Arias Oliva, Kiyoshi Murata, Ana María Lara Palma, pp. 513-525.

Szymoniak, S. (2021b). Security protocols analysis including various time parameters. *Mathematical Biosciences and Engineering,* 18(2): 1136-1153. http://doi.org/10.3934/mbe.2021061

Tokarčíková, E., et al. (2014). Automotive Company's social responsibility in Slovakia, Proceedings of the 24th International Business Information Management Association Conference – Crafting Global Competitive Economies: 2020 Vision Strategic Planning and Smart Implementation, pp. 2118-2127.

Top Priorities for IT: Leadership Vision for 2021 (2020). Gartner, Inc.

7. Values in the Smart Technology Revolution

Adam Poulsen. Brain and Mind Centre, The University of Sydney; Oliver Burmeister. School of Computing, Mathematics and Engineering, Charles Sturt University

EMBEDDING VALUES IN AI BY DESIGN: AN INTEGRATED FRAMEWORK

Xenia Ziouvelou, Vangelis Karkaletsis, Konstantina Giouvanopoulou

AI Politeia Lab, National Centre of Scientific Research "Demokritos" (Greece)

xeniaziouvelou@iit.demokritos.gr; vangelis@iit.demokritos.gr; kgiouvano@iit.demokritos.gr

EXTENDED ABSTRACT

Artificial Intelligence (AI) is evolving rapidly, becoming a key driver for the digital transformation of our economies and societies. Impacting this way the future of humanity, by transforming the lives of individuals and influencing human societies, reshaping patterns of living, working, learning and interacting. However, while AI can create great opportunities by driving economic and social progress, it also presents complex challenges and potential risks. Risks are related to gender-based or other kinds of discrimination and bias (intentional and unintentional), opaque decision-making, intrusion, social harms for individuals and society, loss of liberty, control and autonomy, in addition to the concentration of power in the hands of a few private actors, among others (UNESCO, 2020; EIGE, 2021). Challenges on the other hand, stem from the great uncertainties that are linked with the alignment of AI systems with human values (AI value alignment) from their design to their use (Han et al., 2022); which is of major concern given that the way we model and design AI may affect the values we are able to embed (Gabriel, 2020; Van de Poel, 2020). However, there are other dimensions. The evolution of AI brings about the need to explore deeper the interplay between values and technology design, development, implementation, and use and the role of individuals in realising value sensitive technology; as well as the need explore new values, which are appropriate to protect the rights of the individual in the light of such an evolution (Ziouvelou et al., 2020).

Beyond these anticipated risks, there are increasing concerns over unintended and unanticipated risks with negative, undesirable impacts that may accompany AI technology and its applications. Triggered by these risks, a growing body of ethical AI guidelines and principles, has emerged over the last few years (Hagendorff, 2020, 2022, EU HLEG, 2019; Whittaker et al., 2018; Campolo et al., 2017; Floridi et al., 2018; IEE, 2019; Jobin et al., 2019; among others) aiming to harness the unintended disruptive potential and complex challenges posed by AI. Numerous guidelines have been launched by governments, scientific or industrial communities as well as civil society representatives, aiming to serve as a basis for ethical decision-making in AI design, development, deployment and governance. However, public debate is already saturated by these ethical guidelines. From a macroscopic perspective, this abundance of ethical principles threatens on the one side to overwhelm and confuse and on the other to delay the development of laws, rules and standards that will ensure that AI is socially beneficial (Floridi and Cowls, 2019) or even avoid regulation altogether (Wagner, 2018) in some geographical regions. From a microscopic perspective, the vast majority of these guidelines appear to adopt the 'deontological ethical approach' (Mittelstadt et al., 2019; Hagendorff, 2020), that emphasises duties or rules at an institutional level. At an individual level though, there appears to be a gap, for example in relation to the values, moral and character dispositions of the individuals who create these technologies (at an individual and

company level). Business, government and civil society leaders need to understand the importance of values and ethics in technological development in order to seize the opportunities and address the threats that accompany emerging technologies (seen as sociotechnical systems rather than isolated artifacts (Van de Poel, 2020)), and this implies adopting a conscious perspective on technological development that prioritises society's values (Philbeck et al., 2018). As such, virtue ethics could expand traditional deontological AI ethics and broaden the scope of action (Hagendorff, 2020, Van de Poel, 2020).

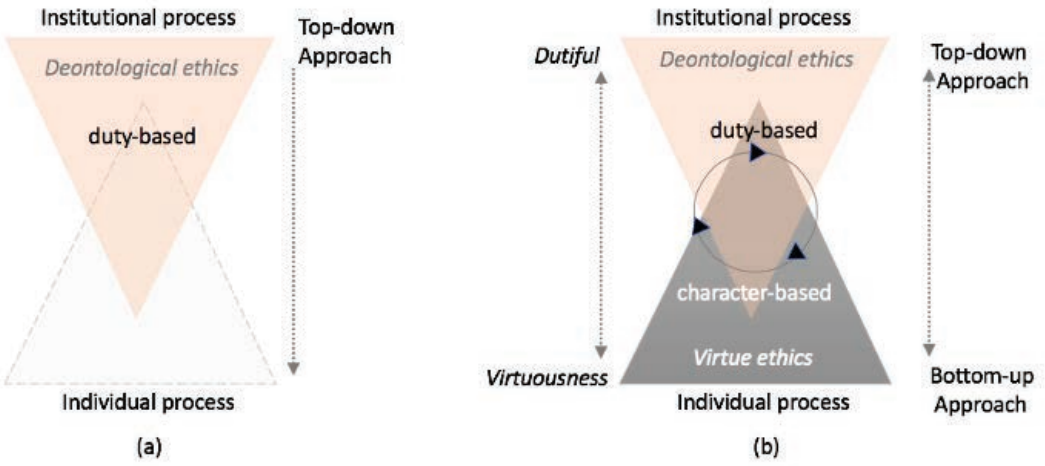
In this paper, we map the existing landscape of systematic scoping studies in the area of ethical guidelines for AI and we examine whether convergence is emerging in relation to the principles, Furthermore, we examine the theoretical parameters underpinning these ethical guidelines, identify gaps and provide a complementary perspective that aims to provide an integrated, multiperspective approach to the discussion about what constitutes 'ethical AI '. We support the view that understanding human values is a crucial step in the development of responsible and trustworthy AI (Han et al., 2022). Our perspective is anchored in the analysis of Hagendorff (2020) and Van de Poel (2020) and the need for a holistic framework for AI ethics that will present a model that augments the traditional, prevalent deontological approach of AI ethics (Mittelstadt et al., 2019; Hagendorff, 2020) (Figure 1a) with an approach oriented towards virtue ethics pertaining values, moral and character dispositions (Van de Poel, 2020). As it is very difficult to predict exactly what kind of consequences future innovations will bring to society, Shannon Vallor (2016) argues for virtue ethics as an appropriate framework for the development of emerging technologies, which, by enabling new forms of behaviour, are expected to influence human values in the future (Steen et al., 2021).

The difference between deontology and virtue ethics is that while the former is based on normative rules with universal validity, the latter examines what constitutes a good person or character. Ethics focuses on the act, while virtue focuses on the actor, on the development of positive characteristics of the actor (Hagendorff, 2020) and is essential if we consider that the values that every individual engineer embraces should be the starting point for responsible and ethical behaviour (Hersh, 2012). Values contribute to evaluation in terms of goodness and badness, while concepts such as duties and rules are used to determine the rightness (or wrongness) of actions (Van de Poel, 2020). The virtuous actor embraces values of goodness, therefore the ethics of virtue is directly related to moral values. As stated by Annas (2011) virtue is a disposition of character, which is not impermanent, to act reliably and virtue requires commitment to values, it involves the orientation of the person to something that the person considers valuable. In an effort to make the implementation of existing AI ethics initiatives successful and effective, the insights of moral psychology should be included, since until now, when talking about AI ethics, the psychological processes that limit the goals and effectiveness of ethics programs are not taken into account (Hagendorff, 2020).

Our motivation is in developing a model that will adopt such an integrated approach to AI ethics that will augment the existing duty-driven approach including principles and rules (i.e., ethical AI code) (Figure 1a - deontological approach) with a virtue-driven approach including values and moral personality traits (i.e., moral/value AI code) (Figure 1b- integrated approach). This model will thus, broaden the scope of action by infusing virtues and ethos in AI ethics. Ethos means ''virtue" (the translation of the Ancient Greek word ἀρετή - ''arete'') in the Aristotelian sense and denotes the internal values that characterise an individual. Aristotle argued that man is by nature zoon politikon, destined to live in an organised political society

and that virtues contribute to living in a polis and promoting the welfare of the people (Steen et al., 2021). The word virtue denotes moral excellence, and it indicates the fundamental qualities that allow people to excel and thus contribute to social well-being. Virtue ethics is a theory that although it cannot guarantee beneficial societal innovation, it can nevertheless be of value in a holistic framework that complements the existing deontological ethics, as illustrated in the figure 1.

Figure 1: (a) Deontological AI ethics approach & (b) Integrated AI ethics framework (Deontological & Virtue ethics).



Considering that artificial systems are not able to understand the notion of human values (Neuhäuser, 2015), lack emotional abilities (Sharkey, 2017) and are human made (Hakli & Mäkelä, 2019), all concern should be about humans involved in designing, developing and deploying AI systems. Given that humans can be considered full moral agents (Dignum, 2018; Hakli & Mäkelä, 2019), virtue ethics suggests that we do not treat AI systems as autonomous, equal to humans, but rather as assistive companions (Maes, 1995; Savulescu & Maslen, 2015; Voinea et al, 2020) as intelligent tools (Balkin, 2017) to serve human needs in a responsible way. Furthermore, just as individuals can demonstrate character, an organization can also embody character, as a collection of individuals (Moore, 2005). Existing research indicates that organizations that demonstrate virtue by exhibiting character, tend to experience positive benefits both internally as well as in the marketplace (Cameron, Bright, and Caza, 2004; Sosik, Gentry, and Chun, 2012; Neubert and Montañez, 2020).

This paper, aims to address the need for a holistic framework for AI ethics by design; a framework that will augment the current prevalent approach with a virtue-driven approach aiming at values, moral and character dispositions of individuals (human embedding values in AI systems (at an individual level and organizational level). To this end it provides an integrated approach to AI ethics aiming to broaden the scope of action by embedding virtues, ethos and values in AI by design. As such there is a need to explore the practical implementation of such a holistic approach and examine how the proposed framework can be implemented so as to foster responsible and trustworthy AI Systems based on an integrated ethics approach, that augments the current deontological approach by using virtue ethics. This will in turn provide

some useful insights into the interplay between virtue ethics and deontological ethics in the context of AI systems and values.

KEYWORDS: Artificial Intelligence, Values, Sociotechnical AI systems, Value embedding, Value-driven AI, AI Ethics Framework.

AKNOWLODGEMENTS: This research is funded by the project AI4EUROPE, Grant Agreement No 101070000 (EU funded project), under the HORIZON.2.4.5 - AI and Robotics.

REFERENCES

Annas, J. (2011). *Intelligent Virtue*. Oxford University.

Balkin, J. M. (2017). The three laws of robotics in the age of big data. *Ohio State Law Journal,* 78(5), 1217–1241.

Cameron, K. S., Bright, D., & Caza, A. (2004). Exploring the relationships between organizational virtuousness and performance. American Behavioral Scientist, 47(6), 766-790.

Campolo, A., Sanfilippo, M. R., Whittaker, M., & Crawford, K. (2017). *AI Now 2017 Report*. AI Now Institute at New York University.

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology, 20*(1), 1–3.

EIGE (2021). Artificial intelligence, platform work and gender equality. *European Institute for Gender Equality.* Luxembourg: Publications Office of the European Union.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., et al. (2018). AI4People - an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28 (4), 689–707.

Floridi, L. and Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*.

Gabriel, I. (2020). Artificial Intelligence, Values and Alignment. *Minds and Machines, 30,* 411-437.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines,* 1-22.

Hagendorff, T. (2022). A Virtue-Based Framework to Support Putting AI Ethics into Practice. *Philos. Technol.* 35, 55.

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*(2), 259–275.

Han, S., Kelly, E., Nikou, S. & Svee, E.Q. (2022). Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI & Soc* 37, 1383–1395.

Hersh, M.A. (2012). Science, Technology and Values: Promoting Ethics and Social Responsibility, *IFAC Proceedings Volumes, 45* (10), 79-84.

HLEG, (2019). A definition of AI: Main capabilities and disciplines, High-Level Expert Group on Artificial Intelligence of the European Commission. *Downloded,* 1, 2019-12.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature*

Moore, G. (2005). Corporate character: Modern virtue ethics and the virtuous corporation. Business Ethics Quarterly, 15(4), 659-685

Maes, P. (1995). Artificial life meets entertainment: Lifelike autonomous agents. *Communications of the ACM, 38*(11), 108–114.

Mittelstadt, B., Russell, C., and Wachter, S. (2019). *Explaining explanations in AI*. In Proceedings of the conference on fairness, accountability, and transparency—FAT* ’19, 1–10.

Neubert, M. J., and Montañez, G. D. (2020). Virtue as a framework for the design and use of artificial intelligence. Business Horizons, 63(2), 195-204.

Neuhäuser, C. (2015). Some Skeptical Remarks Regarding Robot Responsibility and a Way Forward. In C. Misselhorn (Ed.), *Collective Action and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation* (pp. 131–146). Springer.

Philbeck, T., Davis, N. and Engtoft Larsen, A. M. (2018). White Paper. Values, Ethics and Innovation Rethinking Technological Development in the Fourth Industrial Revolution. *World Economic Forum.*

Savulescu, J., & Maslen, H. (2015). Moral Enhancement and Artificial Intelligence: Moral AI? In J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond Artificial Intelligence. The Disappearing Human-Machine Divide* (pp. 79–95). Springer.

Sosik, J. J., Gentry, W. A., & Chun, J. U. (2012). The value of virtue in the upper echelons: A multisource examination of executive character strengths and performance. The Leadership Quarterly, 23(3), 367-382.

Sharkey, A. (2017). Can robots be responsible moral agents? And why should we care? *Connection Science, 29*(3), 210–216.

Steen, M., Sand, M. & Poel, I. (2021). Virtue Ethics for Responsible Innovation. *Business & Professional Ethics Journal.*

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems.*

UNESCO (2020). Artificial Intelligence and Gender Equality, Key findings of UNESCO’s Global Dialogue.

Vallor, S. (2016). Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. New York: Oxford University Press.

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, *30*(3), 385-409.

Voinea, C., Vică, C., Mihailov, E., & Savulescu, J. (2020). The Internet as Cognitive Enhancement. *Science and Engineering Ethics,* 26(4), 2345–2362.

Wagner B. (2018). *Ethics as an escape from regulation: From "ethics-washing" to ethics-shopping?* In Bayamlioglu, E., Baraliuc, I., Janssens, L. A. W. & Hildebrandt, M. (Eds.). Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen (pp. 84-89). Amsterdam: Amsterdam University Press.

Wallach, W. (2004). Artificial Morality: Bounded Rationality, Bounded Morality and Emotions. In I. Smit, G. Lasker and W. Wallach, editors, *Proceedings of the Intersymp 2004 Workshop on Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence,* pp. 1 – 6, Baden-Baden, Germany, IIAS, Windsor, Ontario.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., Schwartz, O. (2018). AI Now report 2018.

Ziouvelou X., V. Karkaletsis, G. Giannakopoulos, A. Nousias, S. Konstantopoulos (2020). Democratising AI: A National Strategy for Greece. NCSR Demokritos http://democratisingai.gr/

SCIENTIFIC RESEARCH IN THE AGE OF LLMs: MORAL CONSIDERATIONS FOR PUBLICATIONS

Michael S. Kirkpatrick

James Madison University (USA)

kirkpams@jmu.edu

EXTENDED ABSTRACT

In the short amount of time since its release, ChatGPT has generated both excitement and trepidation in a variety of fields, such as medicine, writing, and science. ChatGPT is a generative pre-trained transformer (GPT) form of large-language model (LLM). LLMs are a form of artificial intelligence (AI) that apply deep learning techniques to a large body of input text data to train complex neural networks. GPTs use these networks to respond to users' queries with newly generated text that is highly readable and has a high probability of answering correctly based on statistical correlations in the training set of text data. Although the creation and use of LLMs raise questions for a variety of fields (De Angelis, 2022; Gendron et al., 2023; Hughes, 2023; Kolata, 2023), our focus is on how these tools challenge existing policies surrounding publication in scientific research.

The ability to generate new text based on an LLM creates opportunities for improving the quality of scientific literature. Through their ability to synthesize multiple sources of input text, LLMs could be used to summarize existing literature and to generate succinct descriptions of background work. LLMs could also be used to improve the quality of the writing, similar to the way that grammar and spelling correction tools already do. That is, LLMs could be used to rephrase difficult passages or to eliminate redundant text. Furthermore, by improving the quality of the writing, these tools offer an opportunity for increasing equity within scientific fields by reducing or eliminating language barriers that are inherent to the international community (Piatek, 2023; Sakai, 2023).

At the same time, LLMs exacerbate existing or create new challenges for research integrity and the underlying principle of authorship. LLM tools are based on probabilistic links between words as symbols and cannot be considered to demonstrate understanding or mastery of either language or a domain of expertise (Searle, 1980; Bender et al., 2021). As such, LLMs lack the capacity for ensuring the accuracy and semantic nuances of their generated text, which is a significant concern for the accuracy of scientific literature. More to the point, like other generative AI techniques based on deep learning, LLMs exhibit the problem of hallucination (Sajid, 2023; Smith, 2023). As the models do not incorporate axioms of logic or ontologies, the generated text may include plausible but empirically false claims.

Beyond the consequentialist concerns of accuracy, the use of LLMs in scientific literature raise questions about the nature, rights, and responsibilities of authorship. First, if a researcher uses an LLM to generate significant portions of text, it is not clear that the researcher can rightfully claim to be the sole author of the work; the generated text is the product of synthesizing work

written by other people[13]. However, policies regarding authorship are generally based on the assumption that an author is a human who meets certain criteria (ALLEA, 2017; ACM, n.d.; COPE, n.d.; ICMJE, n.d.; Nature, n.d.; ORI, n.d.; Wareham, 2019). Some policies explicitly state that the author is a human, while others have left this aspect implicit. In scientific literature, one of those criteria is acceptance of moral accountability for the work, including the design and implementation of the underlying research. LLMs are not moral agents that can accept this accountability (Johnson, 2006).

Next, LLM tools typically incorporate a feedback loop where the prompts and outputs are used to train future iterations of the underlying model. This feedback may exacerbate the concerns about accuracy, as the text generated during the writing process may contain errors that are later detected during the peer review and editing process. However, the corrections would not be integrated into the model. In addition, depending on how frequently the model is updated, the new results could become inadvertently published (and used without citation) as they are included in others' outputs. Beyond the issues of accuracy and citation, the use of both inputs and outputs by LLMs raise significant questions about copyright protections (Helms & Krieser, 2023), particularly as much of the scientific literature copyrights are owned by the publishers.

Up to this point, we have only been examining the concerns of LLMs by scientific researchers engaged in legitimate practices. However, the public's ability to trust in the findings of research depend on the detection and correction of ethical breaches. These breaches include fraud, such as fabricating results, and plagiarism, including both verbatim copying and intentional paraphrasing of text. These problems are made worse by the presence of paper mills, individuals and organizations that seemingly legitimate scientific work for profit (Nash, 2022). LLMs have the potential to make the problems of paper mills worse, as these tools can greatly expand and automate the production of fake works. Defending scientific integrity relies on detecting these problems, but the detection of work generated by LLMs has mixed results. Although Desaire et al. (2023) report a very high detection rate (99%), their training data set was very selective and unlikely to generalize; they trained their detection tool on 50 papers published in *Science*, an extremely prestigious journal, that is not reflective of the quality of the work contributed to other publishers. OpenAI, the creators of ChatGPT, built a similar detection tool that only detected 26% AI-authored text, while also falsely flagging 9% of human-generated text as AI-authored (Kirchner et al., 2023). These latter results do not bode well for automated detection of LLM-generated works.

The nature of the peer review process in scientific research raises many of these same concerns while introducing others that are distinct from traditional authorship. For instance, a peer reviewer may provide the entire text of a submission as a prompt for an LLM. The LLM could generate a summary (which may be helpful), but the problem of hallucination raises concerns about the accuracy of the review; if the review is determined based on the generated summary rather than the text itself, the publication decision is not necessarily based on the reviewers' expertise-informed judgment. Furthermore, the generated summary might

[13] Arguably, human authors also incorporate others' work in their synthesis and citations. A key difference, though, is that the human authors ostensibly have specialized training that facilitates their judgment in determining relative importance that may not reflect statistical links and frequencies.

incorporate references to the submitting authors' other existing work, thereby eliminating the anonymity of the submission.

The feedback loop created by storing inputs and incorporating them into the model also violates the confidentiality requirements of many publication policies. In common current practices, reviewers are obligated to keep the existence of the submitted work (not just the contents itself) confidential. Systems that store and integrate the inputs into revised models create the possibility that the submitted work, including the authors' identities (for singly anonymous submissions), could be leaked before the work is accepted or published.

Based on these considerations, it is important to consider how publishers should modify existing policies on authorship and peer review. In this talk, we will discuss the merits and concerns of allowing researchers, authors, and reviewers to use LLMs in various ways during the scientific publication process. That is, we will examine how LLMs challenge our existing values surrounding what constitutes authorship and research integrity. In particular, we will focus on the concerns that are unique to scientific publication rather than the more generalized use case of public access to LLMs. Our hope is that publishers will be able to craft policies that balance the advantages and disadvantages of these tools, ensuring accurate recognition of researchers' intellectual contributions and enforcing moral responsibility for the work reported and the integrity of the scientific record.

KEYWORDS: Artificial intelligence, authorship, policy, research integrity, scientific research.

REFERENCES

ALLEA: All European Academies. (2017). The European code of conduct for research integrity. https://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf

ACM: Association for Computing Machinery. (n.d.). ACM policy on authorship. https://www.acm.org/publications/policies/new-acm-policy-on-authorship

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.

COPE: Committee on Publication Ethics. (n.d.). *Authorship*. https://publicationethics.org/resources/discussion-documents/5-what-constitutes-authorship-june-2014

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

Desaire, H., Chua, A. E., Isom, M., Jarosova, R., and Hua, D. (2023). Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science,* 4(6). https://doi.org/10.1016/j.xcrp.2023.101426

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

Helms, S. and Krieser, J. (2023, March). Copyright chaos: Legal implications of generative AI. *Bloomberg Law*. https://www.bloomberglaw.com/external/document/XDDQ1PNK000000/copyrights-professional-perspective-copyright-chaos-legal-implic

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/

ICMJE: International Committee of Medical Journal Editors. (n.d.). Recommendations. Defining the role of authors and contributors. https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

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

Kirchner, J. H., Ahmad, L., Aaronson, S., and Leike, J. (2023, January 31). New AI classifier for indicating AI-written text. *OpenAI*. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

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

Nash, J. (2022, May 9). Paper mills – The dark side of the academic publishing industry. *MDPI Blog*. https://blog.mdpi.com/2022/05/09/paper-mills/

Nature. (n.d.). Authorship. *Nature Portfolio*. https://www.nature.com/nature-portfolio/editorial-policies/authorship

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/

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/

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

Smith, C. (2023, March 13). Hallucinations could blunt ChatGPT's success. *IEEE Spectrum*. https://spectrum.ieee.org/ai-hallucination

ORI: U. S. Department of Health & Human Services Office of Research Integrity. (n.d.). *Authorship*. https://ori.hhs.gov/content/Chapter-9-Authorship-and-Publication-Authorship

Wareham, S. (2019). Editor in Chief (EIC) Manual of the IEEE Computer Society. *IEEE Computer Society*. https://www.computer.org/volunteering/boards-and-committees/resources/eic-manual

ESCAPING THE BENEVOLENT ARTIFICIAL PHYSICIAN: PRIORITIZING CARE ETHICS IN AI-BASED HEALTHCARE

Stacy A. Doore, Ph.D. and Jaime Yockey

Department of Computer Science, Colby College, Waterville, ME, USA

sadoore@colby.edu; mcyock@colby.edu

EXTENDED ABSTRACT

Recent reports estimate that 61 million Americans are affected by a chronic health condition that impacts daily life activities ranging from mobility, cognition, hearing, vision, independent living, and self-care (CDC, 2023). Medicare enrollment is expected to double over the next 15 years, leading to more than 80 million beneficiaries by 2030 (U.S. Centers for Medicare & Medicaid Services, 2021). A large percentage of those with chronic healthcare needs are the result of aging and this has produced a significant shortage in skilled, reliable caretakers in residential facilities and home assistance (Fiorini, 2019; U.S. Department of Labor, 2022). Hospital administrators are hoping that AI powered robotics will provide viable solutions to understaffed medical facilities, home care assistance, and those affected with serious healthcare challenges (Bohr, 2020). However, we argue that for emerging technologies such as care robots to be responsibly integrated into current healthcare sector there needs to be a discussion about how to design and implement based on a care ethics (Gilligan, 1982; Noddings, 2013) framework. This paper begins with a brief discussion of van Wynsberghe's care-centered value sensitive design (CCVSD) framework (2013) with a recommendation to include the principles of justice, transparency, and dignity. It uses a fictional narrative to illustrate why there should be a temporal component in the framework to prevent any shift of foundational values in a system designed and deployed in a specific healthcare context.

Background

Within the field of healthcare robotics, there are several distinctions between the types of healthcare settings and the care tasks they perform. Hospital robots serve a similar function as traditional medical assistants whose primary function is to perform non-critical tasks of monitoring or lifting (Kyrarini et al, 2021). Surgical robots help surgeons with fine precision tasks during surgical procedures. Assistive robots are designed to help patients with activities of daily living (ADL) when there are health conditions such as involuntary movement, limited range of motion, and mobility limitations (Yamazaki et al, 2012). Care robots, the focus of this paper, provide nurses with assistance in more complex patient care tasks, collecting vital health metrics and providing social companionship and interactions for vulnerable patients in hospital or rehabilitation settings (Sharkey & Sharkey, 2010; Vallor, 2011). van Wynsberghe (2013) provides a method for classifying care robots based on three dimensions: application domain (healthcare setting), healthcare use (care practice/tasks), and intended users (giver or receiver of care).

In this paper, we use van Wynsberghe's care-centered value sensitive design (CCVSD) framework (2013) to analyze the application of care robots in a fictional narrative to illustrate

the ways in which there are often conflicting values systems at play with the introduction and use of care robots into healthcare settings that changes over time. Van Wynsberghe centers her framework around Tronto's (2010) fundamental care values *of attentiveness, responsibility, competence*, and *reciprocity* and provides a set of methods to address each one of these concepts during the design phase by examining what these would look like in a specific context with and without the presence of a care robot. While we agree this is a sound place to start, we also believe the framework should be extended to include a temporal component and suggest including additional concepts of related to care ethics such as justice, transparency, and dignity. We now illustrate the rationale for these recommendations based on a narrative about how emerging technologies created through a lens of care ethics and responsible design practices can change over time to produce artificial systems that do not reflect the original values of the designer.

Caring to what end?

In the speculative narrative, *Escaping the Caring Seasons* (Pinkster, 2018), Zora and Anya Stein wrestle with some of humanity's deepest concerns surrounding the future use of AI-based care robots. As a former developer of assistive living facility in a near future setting, Zora designed a rehabilitation hospital that utilized caregiving and diagnostic decision-making AI robots and systems including an AI robot (DOC) to ensure patients were able to return to the comfort of their homes as quickly as possible. As the creator, Zora thought she had embedded a set of values reflecting central premises of care ethics that prioritized a return to independence, relationships with care staff, and communication between systems and staff to build efficiency into an elder care facility. Although a value sensitive design approach (Friedman, 1996) was not explicitly mentioned in the text, the reader is given the impression that this computer scientist was intentional and proactive in the way the system was designed to promote a set of fundamental values based on stakeholder input in the care of patients. However, over time, the hospital that Zora had worked for was acquired by a larger corporation, which made significant changes to the system of care robots to increase automation, reduce on-site administration costs, and maximize profits resulting in a complete loss of autonomy for the patient.

This fictional scenario illustrates the difficulty of maintaining a commitment to an original set of human values in the face of sweeping automation and removal of humans from care roles. Zora is faced with her own loss of autonomy as a caregiver when the imposed restrictions and limitations on Anya prevent her from making decisions for herself and her wife. Seniors in this nursing home are heavily surveilled, their lives are dictated by the care robot's decision-making program through pervasive computer vision sensors and implanted biometric chips that are engaged in the tracking and calculation of their personal health data. Through this system of artificial beneficent care, residents have lost all personal freedom in a sociotechnical system that treats elder care as family burden to be relieved and perhaps even refashioned into a source of popular entertainment. Zora, as the creator of the system of AI powered robot 'caregivers', witnesses the evolution of this system and its shifting of core values and definitions of care over time.

Ethical analysis

Although a technology design and development process may be grounded in a care ethics framework (i.e., CCVSD) that includes all stakeholders to establish its ethical development and use guidelines, the decisions and actions made by humans and AI powered care robots have the potential for harmful consequences for patients and their families. By identifying the moral assumptions in Pinkster's fictional AI care system, we can identify the ways near future systems may fail to account for factors that have significant impacts on an individual's quality of life (or end of life). First, the system has evolved to reduce patient 'wellness' to the state of bodily functions metrics. Second, the lead engineer assumes that embedded values at the design and development phase will remain constant as a system evolves over time. Third, the deployment of care robots for small care tasks to increase human time for meaningful care activities that required the core values of attentiveness, responsibility, competence, and reciprocity (van Wynsberghe, 2013) may lead to the eventual removal of all human caregivers without checks and balances for maintaining its core principles of care ethics.

We suggest there are several other concepts to be considered in all phases of care robotics adoption, beyond the CCVSD four core concepts, that address harms resulting from the erosion of human care giving as illustrated in the example narrative. This move towards efficiency over time is something that van Wynsberghe (2020) also concludes is a potential weakness in her conceptual framework. In response, we propose applying Held's (1995; 2006) "meshing" of justice and care ethics to the CCVSD framework and to stress the need for explicit transparency of decision making by intelligent systems to ensure the fairness in access to emerging care technologies to ensure the system is not based on biased datasets, perpetuating inequities in social systems. In addition, we argue that Ricoer's definition of dignity (1992) that emphasizes community and social relationships in decision making practices should be added to the framework guiding the use of care robots in healthcare settings because ideally the patient should be supported in maintaining their own dignity while those in their social network uphold an attitude of respect to the individual when they are at their most vulnerable (Leget, 2013).

Finally, the addition of a temporal component to the framework moves the values commitment beyond the design and development period to the implementation and auditing stage. The proposal made by Valles-Peris and Domènech (2023), *Caring in the in-between,* calls for practical actions that ensure the consistency of system values over time. This includes the monitoring of relationships and caregiving processes, the engagement of stakeholders to solicit concerns and priorities when making institutional changes, and alleviating fears by instilling freedom of choices in care that are reversible. With these additions, we believe this augmented care ethics framework for the design of emerging healthcare solutions such as care robots may be able to sustain an original set of moral values during the later stages of an intelligent system's deployment and auditing lifecycle.

KEYWORDS: Artificial Intelligence, Care Robots, Ethics of Care, Value Sensitive Design.

REFERENCES

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in Healthcare.* 25-60.

Bureau of Labor Statistics, U.S. Department of Labor, (2022, September) *Occupational Outlook Handbook*, Home Health and Personal Care Aides, Retrieved June 2, 2023 https://www.bls.gov/ooh/healthcare/home-health-aides-and-personal-care-aides.htm

Center for Disease Control and Prevention. (2023). "Disability Impacts All of Us Infographic." *Center for Disease Control and Prevention,* 15 May 2023, Retrieved June 1, 2023 https://www.cdc.gov/ncbddd/disabilityandhealth/infographic-disability-impacts-all.html

Fiorini, L., De Mul, M., Fabbricotti, I., Limosani, R., Vitanza, A., D'Onofrio, GTsui M., Sancarlo D., Giuliani F., Greco A., Guiot D., Senges E., & Cavallo, F. (2021). Assistive robots to improve the independent living of older persons: results from a needs study. *Disability and Rehabilitation: Assistive Technology*, *16*(1), 92-102.

Friedman, B. (1996). Value-sensitive design. *Interactions*, *3*(6), 16-23.

Gilligan, C. (1982). *In a Different Voice*. Cambridge, MA: Harvard University Press.

Held, V. (1995). The meshing of care and justice. *Hypatia*, *10*(2), 128-132.

Held, V. (2006). *The ethics of care: Personal, political, and global*. Oxford university press.

Kyrarini, M., Lygerakis, F., Rajavenkatanarayanan, A., Sevastopoulos, C., Nambiappan, H. R., Chaitanya, K. K., Babu, A. R., Matthew, J. & Makedon, F. (2021). A survey of robots in healthcare. *Technologies*, *9*(1), 8.

Leget, C. (2013). Analyzing dignity: a perspective from the ethics of care. *Medicine, Health Care and Philosophy*, *16*, 945-952.

Noddings, N. (2013). *Caring: A relational approach to ethics and moral education*. University of California Press. Pinkster, S. *Escaping the caring seasons*. (2018) Roush, W. (Ed.). *Twelve Tomorrows*. 157-179. MIT Press.

Ricoeur, P. 1992. *Oneself as Another*. London: University of Chicago Press.

Sharkey, N., & Sharkey, A. (2010). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*.14, 27-40. Tronto, J. C. (2010). Creating caring institutions: Politics, plurality, and purpose. *Ethics and Social Welfare*, 4(2), 158–171.

U.S. Centers for Medicare & Medicaid Services. (2021, December). *National Health Expenditure Projections 2021-2030.* Retrieved June 1, 2023 https://www.cms.gov/files/document/nhe-projections-forecast-summary.pdf

Vallès-Peris, N., & Domènech, M. (2023). Caring in the in-between: a proposal to introduce responsible AI and robotics to healthcare. *AI & Society*, *38*(4), 1685-1695.

Vallor, S. (2011). Carebots and caregivers: Sustaining the ethical ideal of care in the 21st century. *Journal of Philosophy and Technology, 24, 251–268.*

van Wynsberghe, A. (2013). Designing Robots for Care: Care Centered Value-Sensitive Design. *Science and Engineering Ethics*, *19*(2).407-433.

Yamazaki, K., Ueda, R., Nozawa, S., Kojima, M., Okada, K., Matsumoto, K., Ishikawa, M., & Inaba, M. (2012). Home-assistant robot for an aging society. *Proceedings of the IEEE*, *100*(8), 2429- 2441.

DECONSTRUCTING CONTROVERSIAL PREDICTIVE TECHNOLOGIES FOR CHILDREN IN LAW ENFORCEMENT TO IDENTIFY, UNDERSTAND, AND ADDRESS ETHICAL ISSUES

Pinelopi Troullinou, Francesca Trevisan, Elodie Makhoul, Xenia Ziouvelou, Lilian Mitrou , Paola Fratantoni, Dimitris Kyriazanos, Sofia Segkouli

Trilateral Research (Ireland), Trilateral Research (Ireland), National Centre of Scientific Research "Demokritos" (Greece), University of the Aegean (Greece), Zanasi and Partners (Italy), National Centre of Scientific Research "Demokritos" (Greece), Centre for Research and Technology-Hellas (Greece)

Pinelopi.Troullinou@trilateralresearch.com; Francesca.Trevisan@trilateralresearch.com; Elodie.Makhoul@trilateralresearch.com; xeniaziouvelou@iit.demokritos.com; L.mitrou@aegean.gr; paola.fratantoni@zanasi-alessandro.eu; dkyri@iit.demokritos.gr; sofia@iti.gr

EXTENDED ABSTRACT

There is an increasing employment of AI technologies in the civil security sector in the promise of improving efficiency mainly with regard to resource allocation and automatic data analysis. However, the widespread and intrusive uses of AI introduce new challenges, posing threats to fundamental rights and democratic principles (European Parliament, 2021). AI systems may escalate surveillance practices, amplify discriminatory practices and exacerbate pre-existing societal inequalities (e.g., O'neil 2017, Zuboff, 2019). Vulnerable populations, particularly children[14], require special attention in this context (Charisi, 2022; Rahman & Keseru, 2021). To raise awareness on how AI can uphold or undermine children lives and rights, in 2021, UNICEF released a policy guidance on AI for children pinpointing how predictive analytics on children can limit their identities and experience of the world. As more decisions regarding children are being taken with the aid of predictive systems (Hall et al. 2023), it becomes important to understand how these technologies are developed, used, and how they might impact children's rights and lives.

This paper aims at identifying and addressing the ethical and societal impact of predictive technologies designed to identify youth at risk of committing crime. More specifically, the paper discusses how the use of these technologies by law enforcement can result in portraying children as security concerns and the potential negative consequences that may arise from such characterization. We shed light on the risks involved in such practices by analysing the Prokid (Wientjes et al., 2017) case and the controversies surrounding it. Prokid is an identification tool designed for the early detection of young individuals at risk of (re)offending, originally developed by the Gelderland-Midden police force in The Netherlands. Prokid started

[14] According to the United Nations Convention on the Rights of the Child (UNCRC) (1989), children are referred to as those below the age of 18, in addition adolescent and youth refers to those aged 10-19 (WHO, 2014) and 15-24 (United Nations Department of Economic and Social Affairs (UNDESA). For the purposes of this study, we adopt an inclusive definition of 'children' that encompasses all three definitions, considering those 24 years or younger.

to be introduced in 2009 in four pilot regions: Gelderland-Midden, Amsterdam-Amstelland, Brabant Zuidoost and Hollands Midden (Abraham et al., 2011). Over the course of the years, Prokid has gone through several iterations. The initial version, Prokid 12, was designed to assess the risk of criminality of children under 12 years old. A subsequent version of the system has shifted its focus to the age group of 12 to 18 years (Wientjes et al, 2017). The latest version, which is expected to differ substantially from the others, is still under development and will include individuals up to 23 years (Tweede Kamer der Staten-Generaal, 2022). Prokid relies upon existing police data such as reports of children who have come into contact with the police as suspect, victims or witnesses, their addresses, age, gender, the number and types of crimes committed, and additional information about their family and peers. Children data are sorted in a semi-automated way into four risk categories where “red” indicates critical danger, “orange” indicates a problematic situation in regard to the child or their address, “yellow” indicates that a potential risk is developing, and “white” indicates no risk. It was agreed that the police would take follow up actions with children categorized within the red, orange, and yellow categories.

In the analysis of Prokid, we use social controversy mapping as socio-technical tool to unpack the “black box”, understand the functioning of the technology, and evaluate its ethical and societal impact. Mapping and analyzing social controversies is a methodology that draws on the traditions of Science and Technology Studies and Surveillance Studies (Trevisan et al., forthcoming). It consists of deconstructing social controversies as reported in public discourses to identify and map the stakeholders involved in the technology lifecycle and gain a more nuanced understanding of the diverse perspectives, experiences, needs, values, interests, risks and expectations surrounding the technology development. This structured analysis is key to account for the larger social context and needs, uncover common grounds and areas of contentions and ultimately favour human centered approaches to tech development.

We also evaluate compliance with ethical principles on AI for children to inform policy, advocacy, and ethics scrutiny on these practices. By so doing, we flag the diverse factors that need to be considered in order to build systems that are ethical and socially sustainable promoting children’s safety and security minimising potential harms. Furthermore, we specifically evaluate the impact on children’s rights as the potential to interfere with human dignity, right to personality, privacy, and their ability to make decisions about their own lives (right to self-determination).

With this work, we want to emphasize the importance of conducting ethical, societal and fundamental rights impact assessments employing the social controversies deconstruction method to guide technology development and governance models towards promoting the well-being of children and upholding the no-harm principle. Our work makes two unique contributions. Firstly, it offers an evidence-based framework designed to unpack the black box of controversial technologies, support explainability and accountability and understand the dynamics of the diverse discourses and interests. This approach enables a comprehensive analysis of the technology's impact. Secondly, our approach recognizes that technology does not exist in a vacuum, but it interacts with, shapes, and is shaped by society. Therefore, it delves into the broader social understanding and ethical implications of the technology under examination.

KEYWORDS: Children, societal impact assessment, ethics, controversies, responsible AI.

REFERENCES

Assembly, United Nations General. (1989). Convention on the Rights of the Child. *United Nations, Treaty Series*, *1577*(3), 1-23: Retrieved from: http://wunrn.org/reference/pdf/Convention_Rights_Child.PDF.

Abraham M., Buysse W. Loef L., Bram van Dijk B. (2011). Pilots ProKid Signaleringsinstrument 12- geëvalueerd. WODC. http://hdl.handle.net/20.500.12832/1832

Charisi, V., Chaudron, S., Di Gioia, R., Vuorikari, R., Escobar Planas, M., Sanchez Martin, J.I. and Gomez Gutierrez, E. (2022). Artificial Intelligence and the Rights of the Child: Towards an Integrated Agenda for Research and Policy, EUR 31048 EN, Publications Office of the European Union, Luxembourg, 2022, http://hdl.handle.net/20.500.12832/1832

European Parliament, (2021). Artificial Intelligence in Criminal Law and its use by the Police and Judicial Authorities in Criminal Matters. Cain, K. (2012, June 29). European Parliament Resolution of 6 October 2021. Retrieved from: https://www.europarl.europa.eu/doceo/document/TA-9-2021-0405_EN.html

O'neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Hall, S. F., Sage, M., Scott, C. F., & Joseph, K. (2023). A Systematic Review of Sophisticated Predictive and Prescriptive Analytics in Child Welfare: Accuracy, Equity, and Bias. *Child and Adolescent Social Work Journal*, 1-17.

Rahman, Z., and Keseru, J. (2021). *Predictive Analytics for Children: An Assessment of Ethical Considerations, Risks, and Benefits*. UNICEF Office of Research-Innocenti.

Tweede Kamer der Staten-Generaal (2022). Aanhangsel van de Handelingen. Vragen gesteld door de leden der Kamer, met de daarop door de regering gegeven antwoorden 1177. https://www.tweedekamer.nl/kamerstukken/kamervragen/detail?id=2022Z22088&did=2022D52317

Trevisan F., Troullinou P., Fisher E., Kyriazanos D., Bertelli V. (2023) Deconstructing Social Controversies for a Trusted AI Future. Under Review.

UNDESA- United Nations Department of Economic and Social Affairs, “Definition of Youth”, New York, Retrieved from: http://un.org/esa/socdev/documents/youth/fact-sheets/youth-definition.pdf

UNICEF, (2021). Policy guidance on AI for Children. Retrieved from: https://www.unicef.org/globalinsight/reports/policy-guidance-ai-children

WHO (2014), ‘Recognizing Adolescents’, World Health Organization, Geneva, 2014.

Wientjes, J., Delsing, M., Cillessen, A., Janssens, J., & Scholte, R. (2017). Identifying potential offenders on the basis of police records: development and validation of the ProKid risk assessment tool. *Journal of Criminological Research, Policy and Practice*, *3*(4), 249-260.

Zuboff, S., (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Public Affairs.

TRUSTWORTHY AND USEFUL TOOLS FOR MOBILE PHONE EXTRACTION

Anna-Maria Piskopani, Helena Webb, Liz Dowthwaite, Chris Hargreaves, Nicholas FitzRoy Dale, Quentin Stafford-Fraser, Christos Nikolaou, Derek McAuley

University of Nottingham (United Kingdom), University of Nottingham (United Kingdom), University of Nottingham (United Kingdom), HARGS Solutions (United Kingdom), Telemarq (United Kingdom), University of Nottingham (United Kingdom),

anna-maria.piskopani@nottingham.ac.uk; helena.webb@nottingham.ac.uk; liz.dowthwaite@nottingham.ac.uk; chris@hargs.co.uk; nicholas@telemarq.com; quentin@telemarq.com; christos@telemarq.com; derek.mcauley@nottingham.ac.uk

EXTENDED ABSTRACT

Personal mobile phones, particularly smart phones, can be a valuable repository of information about a user's geographical movements, communications behaviours and online browsing history. For that reason, they offer a valuable source of evidence in criminal investigations. In the Mobile Phone Extraction (MPE) process, copies are made of a device belonging to a suspect, complainant or witness in a case and the data extracted is examined by police and others in the criminal justice system over the course of the ongoing investigation. In recent years, the UK criminal justice system has increasingly deployed Mobile Phone Extraction (MPE) and drawn on digital tools to assist with the analysis of data. It is anticipated that these tools will increasingly incorporate state-of-the-art AI techniques (Costantini et al., 2019).

The extraction and use of personal digital data can provide important evidence to secure criminal convictions but also raises significant ethical and responsibility concerns. Whilst the MPE process can elicit valuable evidence, it is often very lengthy and inefficient. Analysing data from digital devices is resource and time intensive. Forensic examinations can take an extensive amount of time to complete, creating case backlogs (HM Crown Prosecution Service Inspectorate, 2019). Many of the commercially available digital forensics tools are expensive to purchase, limited in usability, and quickly out of date. A lack of accepted standards and validation procedures means that the accuracy of tools cannot necessarily be guaranteed (Horsman, 2019). In addition, extracting and reviewing the entire dataset from a phone may cause distress to the phone owner and be regarded as an intrusion of privacy (Big Brother Watch 2019; Centre for Women's Justice, 2019). Where the extraction is conducted without consideration of what is necessary and proportionate, this raises significant risks of violating data subjects' rights (ICO, 2020; ICO, 2021).

These concerns have culminated in a crisis of trust and practice over MPE, creating a situation in which individuals may decline to hand over their devices to police or even decline to report their experience of crime. In 2020 the UK Information Commissioner's Office argued for action to assure the legality of MPE and for privacy by design to be embedded into digital forensics tools (ICO, 2020). There is clearly an urgent need for MPE tools, and wider practices around their use, that are both useful and trustworthy and that address existing concerns in order to resolve the current crisis of trust and practice.

In this position paper we outline an ongoing research project (Trustworthy and Useful Tools for Mobile Phone Extraction) that identifies opportunities for responsible MPE in the criminal justice system in the UK. The project includes the development of a privacy-preserving digital forensics platform. The RIME (Responsible Investigation of Mobile Environments) platform is designed to expose a subset of the contents of a phone for investigation rather than the entire dataset. This returns a relevant and manageable volume of data and also protects the phone owner's privacy. A further privacy-preserving mechanism is a pseudonymisation feature that replaces real names and phone numbers with autogenerated (but indexed) alternatives. As the project continues, RIME will embed visualisation features to support usability. RIME's modular nature allows the wider community to build plugins to analyse specific apps within its generic framework, and to extend its visualisation and analysis tools to meet specific investigative needs. As an open-source tool, it is also available for assessment and inspection by stakeholders across the MPE process – an important responsibility mechanism.

Our project also involves the conduct of interlinked research activities examining different dimensions of trustworthiness and usefulness in relation to MPE. These research activities draw on computer science, digital forensics, data security, social science, human computer interaction and law to forge a broad and inclusive understanding. The results of these activities inform the ongoing development of RIME and also serve to highlight opportunities for responsibility in MPE more generally.

Firstly, we explore the regulatory framework, with particular attention to data protection issues raised by MPE and MPE tools, and recent legal attempts to mitigate these risks. We conduct stakeholder engagement activities to identify perspectives across the MPE process. These involve participants from

the police, the legal profession, victim support organisations, privacy campaign organisations, and academics specialising in law and technology. We combine the results of these activities with case studies from the existing literature, a review of suggested frameworks for privacy preserving mobile forensics (e.g. Heo et al., 2022; Hyder et al., 2022) and the review of existing standards for digital forensics (e.g. ISO17025 and CASE) to identify ways in which the MPE process can be supported to become more efficient whilst also embedding privacy by design and safeguarding considerations into MPE tools.

We anticipate that our project outcomes will serve multiple purposes by highlighting opportunities for responsibility in MPE and developing a digital forensics tool that is accessible, useful and trustworthy. In particular, we seek to demonstrate how the rights of phone owners can be better protected, and the efficiency of investigation processes can be enhanced.

KEYWORDS: mobile phone extraction, digital forensics, data protection, privacy, criminal justice.

REFERENCES

Big Brother Watch (2019) Digital Strip Searches: The Police's data investigations of victims. Big Brother Watch, July 2019. Retrieved from https://bigbrotherwatch.org.uk/wp-content/uploads/2019/07/Digital-Strip-Searches-Final.pdf

Centre for Women's Justice (2020) Stop the Digital Strip Search of Rape Victims Like Me. Centre for Women's Justice, March 2020. Retrieved from: https://www.centreforwomensjustice.org.uk/new-blog-1/2020/3/13/stop-digital-strip-search

Costantini, S., De Gasperis, G., & Olivieri, R. (2019). Digital forensics and investigations meet artificial intelligence. *Annals of Mathematics and Artificial Intelligence*, 86(1), 193–229. https://doi.org/10.1007/s10472-019-09632-y

Heo, O., Koo, H.J. and Kwon, H.Y., 2022, June. A Study on Privacy Protection of Mobile Evidence in Relation to Criminal Investigative Procedures. In *DG. O 2022: The 23rd Annual International Conference on Digital Government Research* (pp. 346-355).

HM Crown Prosecution Service Inspectorate (2019) *Rape Inspection 2019*. HMCPSI, Publication Number CP001:1267, December 2019. Retrieved from https://www.justiceinspectorates.gov.uk/hmcpsi/wp-content/uploads/sites/3/2019/12/Rape-inspection-2019-1.pdf

Horsman, G. (2019). Tool testing and reliability issues in the field of digital forensics. *Digital Investigation*, 28, 163–175. https://doi.org/10.1016/j.diin.2019.01.009

Hyder, M.F., Arshad, S., Arfeen, A. and Fatima, T., 2022. Privacy preserving mobile forensic framework using role-based access control and cryptography. *Concurrency and Computation: Practice and Experience*, *34*(23), p.e7178.

ICO (2020) Mobile Phone Data Extraction by Police Forces in England and Wales. Information Commissioner's Office, June 2020. Retrieved from: https://ico.org.uk/media/about-the-ico/documents/2617838/ico-report-on-mpe-in-england-and-wales-v1_1.pdf

ICO (2021) Mobile Phone Data Extraction by Police Forces in England and Wales. An update on our findings. Information Commissioner's Office, June 2021. Retrieved from: https://ico.org.uk/media/about-the-ico/documents/2620093/ico-investigation-mpe-england-wales-202106.pdf

THE CHALLENGE OF CO-CREATION: HOW TO CONNECT TECHNOLOGIES AND COMMUNITIES IN AN ETHICAL WAY

Kristina Khutsishvili, Neeltje Pavicic, Machteld Combé

University of Amsterdam (the Netherlands), City of Amsterdam (the Netherlands), City of Amsterdam (the Netherlands)

K.Khutsishvili@uva.nl; N.Pavicic@amsterdam.nl; M.Combe@amsterdam.nl

EXTENDED ABSTRACT

The contribution aims to reflect on ongoing experiences of connecting the members of technological world with vulnerable and marginalised communities inside the framework of CommuniCity Horizon Europe project.[15] The project draws on three rounds of open calls starting in cities of Porto, Amsterdam and Helsinki and then 'replicated' in other European cities during the project timeframe of 3 years. In the beginning of the open call rounds, hosting cities announce societal challenges to which the pilot proposals need to respond. The selected by independent jury pilots aim to develop technological solutions tailored to the specific needs of local communities together with the members of those communities, by means of co-creation.[16] The overall aspiration of the project is to accumulate the experiences and learnings on co-creation with disadvantaged groups, to come up with scalable practices and solutions, with the possibility to use the guidelines for successful open call and piloting processes as well as technical components and tools to replicate solutions in other cities and communities.

At the time of submitting the initial proposal to the ETHICOMP2024, the midterm meetings with the pilot teams are being run and the co-creation sessions with the targeted groups, meaning the members of communities, are being held. The theoretical, ethics-related question that derives from the related empirical observations and exceeds them – seems to be extremely important for future work aimed at the very same direction. The questions addressed by the contributions are:

What is meant by co-creation when we speak about co-creation of the technological solution with the communities? Where can we draw the line between co-creating with the community and 'testing' the solution on community? How to develop tech solutions for and with marginalized groups without harming or disappointing them?

These questions are answered by means of reflection on more 'practical' issues, among which are: does creating technology with and for the communities imply that the solution is better to be 'built' from scratch? is 'feeding' the application/platform/technological solution with the data coming from the communities, especially vulnerable and marginalised communities, can be viewed as an exercise of co-creation? can the potential positive externality of making

[15] The website of the project: https://communicity-project.eu.

[16] The overview of the pilots that have been selected by the independent jury as a result of the first open call is given here: https://communicity-project.eu/2023/06/22/piloting-teams-announced/.

the solution less biased through engagement with the members of 'target' communities be seen as a balancing act?

In our contribution, the analysis is drawn on empirical observations of piloting processes and related activities. The conceptual part, in turn, starts from the central notions. Aiming to bring together the world of technologies and vulnerable and marginalised communities, we inevitably face questions on the very essence of the conditions of vulnerability and marginalisation. On the grant proposal stage, the wording 'hard to reach' had been used while describing the communities at the centre of attention. Later we decided to abandon such a phrasing, on the grounds of the points raised such as: "Nobody is per se hard to reach"; "Such a terminology suggests that the municipalities are 'lazy' to reach the groups."

The conditions of vulnerability and marginalisation have different focuses, with vulnerability being the 'inner' condition, an inward situation, while the condition of marginalisation implies being an 'object' of the process of marginalisation, being on the 'receiving side' of the external process, in contrast with the 'inner' condition. The word 'disadvantaged', in turn, is used as a comprehensive notion including the notions focused on different aspects and conditions belonging to the 'disadvantaged' condition. We may outline that the words 'disadvantaged', 'vulnerable', and 'marginalised' vary with respect to their sensitivity, with 'disadvantaged' being more neutral. Both conditions of vulnerability and marginalisation relate to ethical considerations: the 'do no harm' normative principle is supplemented by thoughts on the desirable 'empowering' effect relevant for both conditions, being it the 'inner' condition of vulnerability or the vectored toward the person or group, directed from the 'outside', condition of marginalisation.

It is important to emphasise the ethical complexity of the goal set for the project determined not only by communities and their needs being in the centre of attention. Grants distributed to the winning teams are limited in their amount, as is the piloting period. While in cases of internal resources available, the pilot hosts have an incentive to proceed with projects further, such a scenario is not certain and depends on many factors. Keeping in mind these limitation factors, the main goal of the project is to acquire learnings that will enable replication in other cities and communities. The learnings in their broader sense then include not only successes but failures. At the same time, with the communities being at the centre of the processes, not all 'failures' may be desirable, if we may formulate it in this way. Some failures, the failures potentially having a negative impact to the communities involved, need to be minimised. A replication goal embedded in the project, with so-called replicator cities joining second and third open calls and piloting rounds, as well as piloting funding and timeframe limitations and the focus on learnings derived from the processes, all indicate the experimental nature of the project. The dualism of experimentation with the focus on disadvantaged communities brings unprecedented analytical and research possibilities but also the stress on responsibility and the ethical component.

Running pilots with such groups and involving them in creating and developing solutions can generate valuable technological innovations. Consequently, the long-term value of the pilots and experiments for other people who have similar needs, is quite clear. However, the individuals who take part in the project often may not profit from the results of the pilots themselves, or even if they do, it can take a long time. A few relevant examples from the project: a company ran a pilot for a technology increasing the autonomy of the elderly, the elderly people involved liked the technology, but when the piloting period is over it will be

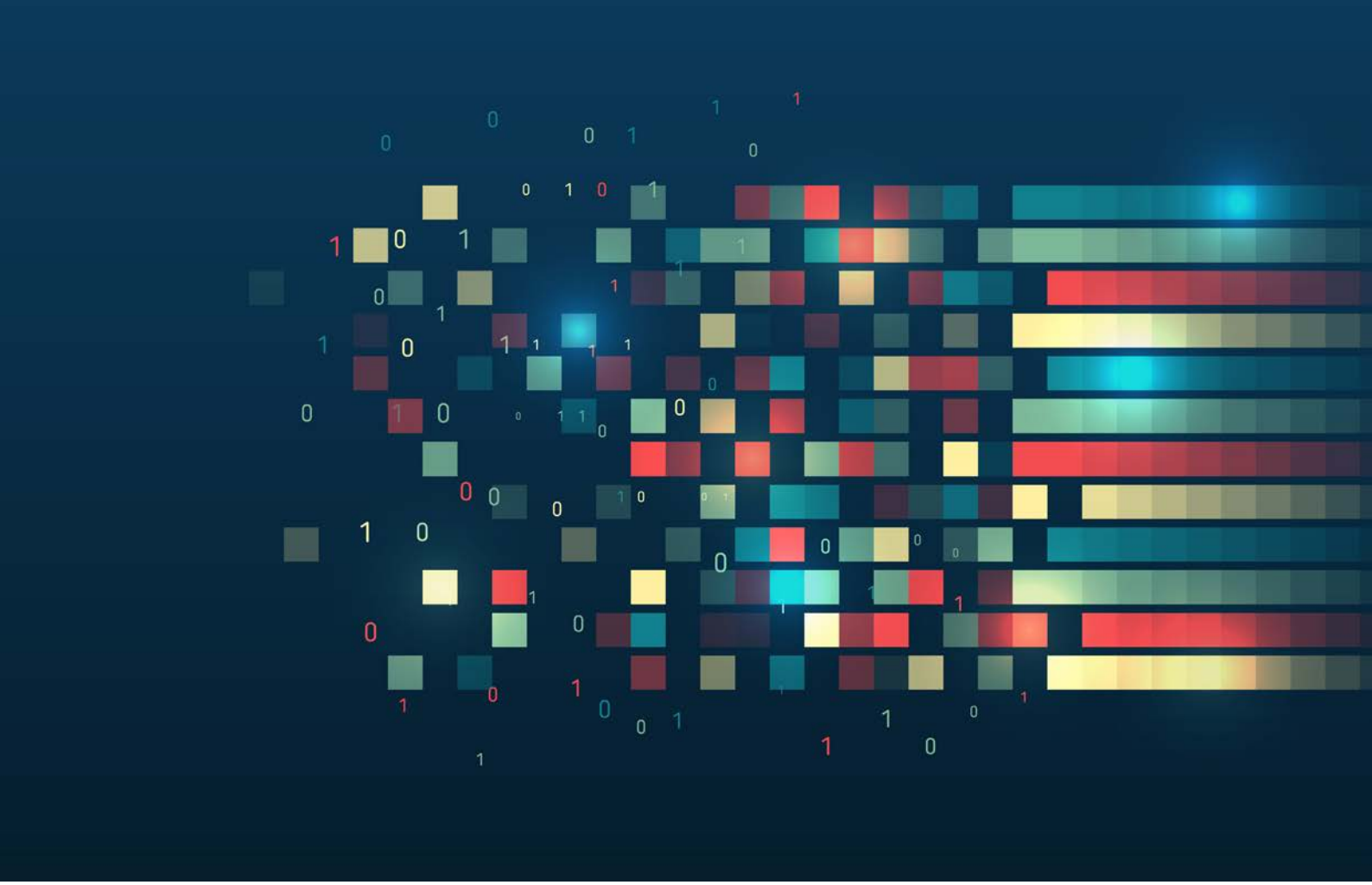
taken away from them as this was just a pilot to learn from. Another example is a platform that is further developed with a group of youngsters who have been in contact with the law, their input improved the platform but the platform turned out to be too expensive for the involved department of the piloting city to be actually implemented. In general, the common opinion is that when you start a pilot you have to invest in expectations management and explain to the people taking part that the objective is to gather knowledge and experiment and that they should not expect that they will be able to use the solution once the piloting is finished. Members of disadvantaged communities, in turn, may be in a position of need so such a 'warning' is not registered very well, and they still hope for a real solution and may be disillusioned at the end of the pilot. At the same time, the evidence suggests also positive, identitarian consequences of engagement, with community members reflecting on their motivation using phrasing such as "helping people, working together towards a higher goal".

Co-creation activities aimed at the communities in question do not make the piloting processes easier, quite the opposite. Yet, the opportunities for experimentation and learning accumulation enabled by such design are extremely valuable. To facilitate and conduct the related activities of community engagement activities in an ethical way, it is necessary to keep in mind the 'do no harm' principle, the power imbalance including the imbalance of professional, technological subject-related knowledge, and the general condition of belonging to a disadvantaged community. The balancing act as well as the risk of harm mitigation act may be exercised by providing clear and honest communication including the communication on general aims and limitations of the project and particular pilot, encouraging the dialogue on equal terms, aimed at 'de-objectivization' of community and its members and empowering the members of communities from the very beginning of the engagement.

KEYWORDS: Ethics, technology, AI, city, community, vulnerability, co-creation.

REFERENCES

The website of CommuniCity Horizon Europe Project: https://communicity-project.eu.



We are living in a smart world, where everything looks smart. The use of the term "smart" is a buzzword. Technology becomes the backbone of virtually every aspect of our life: work, relations, health, education, leisure, ... In ETHICOMP 2024 International Conference, we wonder if the current state of technological revolution is truly smart. What does "smart" really mean in digital contexts, and what should "smart" signify? If technology becomes smart, what is the impact on computer ethics and digital ethics? It is unquestionable that any smart technology must be ethical in both its development and its uses, but cases of unethical practices in the digital world, which are often hidden by technological determinism, are increasing. The analysis of unethical practices and the search for solutions to create a digitally healthy society is the theme of our next conference: digital ethics should lead the smart revolution, but is it doing so? What should we do to locate digital ethics at the core of the smart revolution? What are the ethical requirements for smart in order that it will sustain societal welfare?

UNIVERSIDAD DE LA RIOJA

Cátedras Telefónica

UNIVERSIDAD COMPLUTENSE MADRID

UNIVERSITAT DE BARCELONA

UNIVERSITAT ROVIRA i VIRGILI

Logroño