Embedding AI in society: ethics, policy, governance, and impacts

Article *in* AI & SOCIETY · June 2023

DOI: 10.1007/s00146-023-01704-2

6 authors, including:

Veljko Dubljevic
North Carolina State University
145 PUBLICATIONS **1,831** CITATIONS

SEE PROFILE

William Bauer
North Carolina State University
18 PUBLICATIONS **167** CITATIONS

SEE PROFILE

George F. List
North Carolina State University
127 PUBLICATIONS **2,525** CITATIONS

SEE PROFILE

AI & SOCIETY
https://doi.org/10.1007/s00146-023-01704-2

EDITORIAL

Embedding AI in society: ethics, policy, governance, and impacts

Michael Pflanzer[1,2] · Veljko Dubljević[2,3] · William A. Bauer[3] · Darby Orcutt[2,4] · George List[5] · Munindar P. Singh[6]

Received: 29 May 2023 / Accepted: 31 May 2023

1 Introduction

Artificial Intelligence (AI) is fast becoming a ubiquitous part of our lives. From autonomous vehicles and AI personal assistants to computer-assisted surgery and automated trading systems, we are becoming increasingly reliant upon AI to facilitate decision-making and manage our personal and professional lives. In all these cases, AI promises improvements in efficiency, productivity, and/or safety. However, AI does not simply, automatically, and seamlessly integrate into our daily lives and social institutions. Rather, it directly reshapes social, cultural, and economic structures and affects the lives of individual citizens in profound and often tacit, unpredictable, or morally questionable ways. AI systems have the potential to reweave or even disrupt our socioeconomic fabric, impacting not just our productivity and safety but also our autonomy and dignity.

In recognition of AI's profound potential for benefit and harm, we organized a multidisciplinary symposium that sought a deeper and more holistic understanding of how AI does and should shape societal activity. This Special Issue on Embedding AI in Society encapsulates those findings. Some of these papers were first presented at the symposium before being submitted for the Special Issue, while others were submitted directly for consideration.

Our focus was on research on the social, political, and ethical dimensions of AI. The articles selected revolve around four common themes, some with a conceptual focus and some with an empirical focus:

(1) the relationship between humans and AI,
(2) the ethical principles of AI,
(3) the ethical issues related to the implementation and use of AI, and
(4) the value of domestic and international regulatory frameworks for AI.

Written by a diverse set of scholars, the articles discuss AI across multiple domains, including autonomous vehicles, healthcare robots, policing algorithms, and AI personal assistants.

Below is a synopsis of the contributions in the assembled articles, with an objective of helping readers identify conceptual connections among the arguments presented.

2 Human–AI relationships

The first focus is on human–AI relationships. AI has the potential to transform various aspects of our lives including social interactions, work, and personal identity (cf. Pflanzer et al. 2022). As AI becomes ubiquitous and more advanced, it will undoubtedly alter relationships among humans, humans with AI, and between humans and their environments. The interactions among humans is of particular importance, as AI may obviate many needs to interact. Undoubtedly, it will change our personal and professional lives by automating many jobs and changing the nature of our interactions. One article examines the vocational implications of AI technology in the field of medicine (Kempt et al. 2022). The authors illuminate potential disagreements between physicians and AI-based decision support systems,

✉ Veljko Dubljević
veljko_dubljevic@ncsu.edu

1 Communication, Rhetoric and Digital Media Program, North Carolina State University, Raleigh, NC, USA

2 Science, Technology and Society Program, North Carolina State University, Raleigh, NC, USA

3 Department of Philosophy and Religious Studies, North Carolina State University, Raleigh, NC, USA

4 NC State University Libraries, North Carolina State University, Raleigh, NC, USA

5 Department of Civil, Construction and Environmental Engineering, North Carolina State University, Raleigh, NC, USA

6 Department of Computer Science, North Carolina State University, Raleigh, NC, USA

Published online: 24 June 2023

and also discuss moral responsibility within a more automated clinical work environment.

Personal interaction is the focus of an important subtheme. AI is changing how we communicate and interact with each other through social media, chatbots, and other digital technologies. As AI becomes more sophisticated, it will further shape how we relate, raising important questions about social norms, privacy, and human connections. Grandinetti (2021) examines transparency in the context of Facebook and TikTok to show how AI is becoming embedded. Grandinetti sees AI as a material-discursive apparatus, in that it creates implicit teams of humans and machines that rely on discursive techniques and changing material arrangements. Haque et al. (2023) also examine the effects of AI on social networks by designing a social simulation to analyze the effects of content sharing on polarization and user satisfaction. They conclude that (1) user tolerance slows down polarization but lowers satisfaction; (2) higher selective exposure leads to higher polarization and lower user reach; and (3) both higher tolerance and high exposure lead to a more homophilic social network.

AI also has the potential to shape personal identities—to change the way we see ourselves as well as our place in society. For example, AI may enhance our cognitive abilities, alter our memories, or create entirely new forms of augmented intelligence. These possibilities raise important questions about what it means to be human and how human characteristics should be defined. Munn and Weijers (2022) explore the notion that AI chatbots may become digital friends, asserting that many people see these chatbots as their *best* friends. The authors examine the implications of discontinuing access or removing features. They conclude that lawmakers should endeavor to legally protect people from the adverse effects of losing their "digital friends." The relationships between humans and AI will continue to have significant impacts on our personal and social lives. For example, as is now well known, AI-powered decision-making systems can perpetuate bias and discrimination or even manipulate people's behavior. AI systems are increasingly operating autonomously, outside the sphere of direct human oversight. The authors assert that we should be cognizant of these impacts and work to shape AI in ways that helps it align with broadly shared human values and promote the well-being of all citizens.

3 The ethical principles of AI

It is commonly asserted (see, e.g., Noble and Dubljević 2022) that we should consider the societal impact of AI implementation in the context of ethical values. Unsurprisingly, ethical principles of AI are a major theme for many of the authors whose work appears here. For example, Slota et al. (2021) conducted interviews with 26 stakeholders to explore the challenges of AI, including the distribution of agent empowerment and the difficulty of creating accountable systems. They propose the creation of accountable sociotechnical systems (cf. Chopra and Singh 2021) that can be challenged, interrogated, and adjusted to prevent unjust risk. Such systems must not only demonstrate agency, but also be transparent. Andrada et al. (2022) attempt to classify forms of transparency in human–technology interactions. They conclude that all forms of transparency should be considered when designing ethical AI systems.

AI also affects fairness. Like transparency, AI's social impact is an important ethical principle to consider and debate. Maas (2022) examines fairness by looking at power asymmetry among stakeholders—including those who shape AI (such as developers) and those who are affected by it (such as users). Maas bases the analysis on the concept of domination and suggests that external auditing and design-for-value approaches (see, e.g., Liscio et al. 2022) can mitigate the adverse effects of asymmetrical power. For example, in the context of transportation, Gaio and Cugurullo (2022) suggest that society should prioritize mobility justice over policies that focus on single transportation modes. They argue their case using societal goals of proximal cities and urban containment. Finally, Yazdanpanah et al. (2022) suggest a comprehensive research agenda to support the advancement of responsible AI. Like some of our prior work (see, e.g., Singh 2022), they argue that the rollout of any autonomous system should not only follow a demonstration of trustworthiness, but also an explanation of how the AI responsibly satisfies a societal need.

The normative nature of AI is also a frequent topic, as it can make decisions which affect humans in the real world. Several authors argue there is a need to develop and evaluate ethical theories about what makes actions morally right or wrong. Normative ethics is situated between metaethics (which asks whether ethical decision-making is cognitive or non-cognitive, whether moral values objective or subjective, and so on) and applied ethics (which asks, for example, whether abortion is ethically permissible, or whether war is ever justified). In some of the earliest discussions of AI ethics, a primary question asked was which kind of ethical theory should be implemented (so-called 'machine ethics'). As some of us have noted in earlier work (see, e.g., Dubljević 2020 and Coin and Dubljević 2022), it is important to ask whether AI should make moral decisions like a consequentialist, a deontologist, a virtue theorist, or more simply on case-by-case basis. These questions continue as subjects of intense debate, which some of the articles address.

For example, Begley (2023) argues that normative ethics should not be the beginning of philosophical investigations. He suggests a non-methodological approach which proceeds on a case-by-case basis. The key to this approach

is to ask ethical questions which are meant to spur investigations. Stenseke (2021) outlines a method of implementing ethics in machines that follows the core features of virtue ethics. Following a critical evaluation of the challenges of extending virtue ethics beyond theory into implementation, Stenseke proposes a solution that includes moral functionalism, bottom-up learning, and eudaimonic reward. They conclude by presenting a comprehensive framework for developing artificial virtuous agents and discussing how to implement them into moral environments. Kaluža (2022) explains the shortfalls in addressing the challenge of the "filter bubble" and suggests that the better adaptation method would be habitual. Kaluža then shows that, although habitual adaptation of algorithmic personalization is in contrast with society as it stands, it could explain the adoption and stubbornness to stick to certain kinds of information within an isolated social chamber. Haque et al. (2023) address similar challenges but from the perspective of social simulation.

4 Ethical implementation of AI

The third theme we identify is ethical implementation. It is similar to the second but distinct in that it focuses on specific contexts in which the ethical principles discussed previously are salient. While these articles contextualize their discourse with established ethical principles, their focus on specific domains ties these articles together. Regarding warfare applications, for example, Omotoyinbo (2022) argues that smart robotic soldiers would help address moral challenges of warfare. However, the author also remarks that this approach is extreme and that there are inherent issues with replacing humans with robots (e.g., ethical principles such as responsibility and accountability).

Chatbots are another example. They have become a major focus in recent discussions of AI ethics. Chat-GPT and similar applications are creating major impacts. Fyfe (2022) examines their use in education. Fyfe asked students to use OpenAI's GPT-2 for a final writing assignment and to later reflect upon the ethical implications of utilizing AI chatbots while writing. He used these student reflections to consider the larger conversation of the ethical use of AI and language models. Inasmuch as Chat-GPT and similar applications have risen in visibility since this special issue was developed, we are eagerly following the emergent conversation about the ethical implementation and regulation of such technologies.

Many of the manuscripts also focus on the disciplines that are most likely to be drawn into the ethical AI debate. Examples are the regulatory and legal debates about the implementation and use of AI applications. Novelli (2022) justifies a claim that AI entities should be given personhood, demonstrating the potential liability and harmful behavioral concerns that might arise if this is not done. He also discusses other potential legal ramifications of personhood like contracts and lawsuits. Similarly, Jenkins et al. (2022) lay out a two-phase framework for assessing the consequences, good and bad, of AI systems by examining their use in journalism, criminal justice, and the law. They argue that the legal system is likely to provide much commentary on ethical principles such as justice, fairness, accountability, and responsibility.

5 Calls for domestic and international regulation of AI

Anthropologists, sociologists, and other researchers in related disciplines also focus on these principles and use them to rally policymakers and regulators to responsibly consider the ethical dimensions of AI. Freitas and colleagues (2022), for example, explore the use of AI to characterize neighborhood income and socioeconomic characteristics in urban environments. They suggest that policymakers and politicians could be using such models to justify the benefits of gentrification. They cite the ability of these models to examine the effects of economic and public health crises insofar as urban spaces are concerned. The authors lay out some of the benefits of integrating AI models into the decision-making process. They assert that AI-based models will enable scholars in the humanities to better articulate research questions.

Democratization of AI is another important subtheme. Some articles address questions about how to implement transparency, fairness, justice, and responsibility, with debates over AI's social impacts arguing for the democratization of AI to better realize those principles. Himmelreich (2022) examines the call to democratize AI, arguing that it does not meet legitimization demands, introduces redundancies in the governance of AI, and causes various injustices. However, Himmelreich proposes a better way to democratize AI that avoids the identified problems: Rather than merely focus on fostering increased participation, efforts to facilitate democratization should instead enrich and improve existing infrastructure.

Several of the articles examine the impact of using AI for international affairs. Borsci et al. (2022) examine the European Union Commission's whitepaper on AI and identify two issues with implementation: (1) lack of EU vision and methods to drive decisions at lower levels of government, and (2) support for the diffusion of AI in society. They suggest that research, encouraged by regulators, should seek to see how socioeconomic differences could lead to a fractured AI market. Bisconti et al. (2022) explore ways to maximize the benefits of interdisciplinary cooperation in AI research

groups and explain that this is a temporal urgency given the "AI Act" and other initiatives being undertaken by the EU Commission. They conclude by identifying law enforcement, criminal justice, and social robotics as relevant fields that may benefit from their methodology. Hassan (2022), on the other hand, explores AI governance and regulation gaps in the context of African nations. He demonstrates the existence of Euro-American biases within AI ethics scholarship and identifies a need to consider *non*-Eurocentric perspectives regarding AI ethics, specifically advocating ethical principles from an African perspective.

6 Concluding remarks

The landscape of AI ethics, and more broadly AI in society, is vast in methods, questions, and proposals. The papers in this special issue collection reflect this vastness while raising as many questions as they answer. We certainly encourage more work on the highlighted themes. That said, going forward we also suggest that researchers focus more attention on political power and policy making processes (cf. Dubljević 2019), as well as the possibilities of shared values across pluralistic societies. The questions which need to be explored in the future include: What are the commonalities and differences? And, should we work towards greater moral unity or does the friction of disunity generate new and better ideas? The challenges of trust (see e.g., Singh and Singh 2023), are another area which needs more sustained scholarship. Finally, we see room for more metaethical debate in the discussion of AI ethics, asking: How much hope or trust should we have that we can solve AI ethical problems? How can moral values be implemented and realized in artifact–human relations? And what methods of investigation and ways of knowing are likely to resolve value conflicts?

Acknowledgements The work on this special issue was supported by the award Rabb Science & Society grant 'Embedding AI in Society' 2020. Special thanks to Nora Edgren for her preparatory and early editorial work.

References

Andrada G, Clowes RW, Smart PR (2022) Varieties of transparency: exploring agency within AI systems. AI Soc. https://doi.org/10.1007/s00146-021-01326-6

Begley K (2023) Beta-testing the ethics plugin. AI Soc. https://doi.org/10.1007/s00146-023-01630-3

Bisconti P, Orsitto D, Fedorczyk F et al (2022) Maximizing team synergy in AI-related interdisciplinary groups: an interdisciplinary-by-design iterative methodology. AI Soc. https://doi.org/10.1007/s00146-022-01518-8

Borsci S, Lehtola VV, Nex F et al (2022) Embedding artificial intelligence in society: looking beyond the EU AI master plan using the culture cycle. AI Soc. https://doi.org/10.1007/s00146-021-01383-x

Chopra AK, Singh MP (2021) Accountability as a foundation for requirements in sociotechnical systems. IEEE Internet Comput (IC) 25(6):33–41. https://doi.org/10.1109/MIC.2021.3106835

Coin A, Dubljević V (2022) Using algorithms to make ethical judgments: METHAD vs. the ADC model. Am J Bioeth 22(7):41–43

Dubljević V (2019) Neuroethics, justice and autonomy: public reason in the cognitive enhancement debate. Springer, Heidelberg Germany

Dubljević V (2020) Toward implementing the ADC model of moral judgment in autonomous vehicles. Sci Eng Ethics. https://doi.org/10.1007/s11948-020-00242-0

Freitas F, Berreth T, Chen Y, Jhala A (2022) Characterizing the perception of urban spaces from visual analytics of street-level imagery. AI Soc. https://doi.org/10.1007/s00146-022-01592-y

Fyfe P (2022) How to cheat on your final paper: assigning AI for student writing. AI Soc. https://doi.org/10.1007/s00146-022-01397-z

Gaio A, Cugurullo F (2022) Cyclists and autonomous vehicles at odds. AI Soc. https://doi.org/10.1007/s00146-022-01538-4

Grandinetti J (2021) Examining embedded apparatuses of AI in Facebook and TikTok. AI Soc. https://doi.org/10.1007/s00146-021-01270-5

Haque A, Ajmeri N, Singh MP (2023) Understanding dynamics of polarization via multiagent social simulation. AI Soc. https://doi.org/10.1007/s00146-022-01626-5

Hassan Y (2022) Governing algorithms in the south: AI and sustainable development in Africa. AI Soc. https://doi.org/10.1007/s00146-022-01527-7

Himmelreich J (2022) Against "Democratizing AI." AI Soc. https://doi.org/10.1007/s00146-021-01357-z

Jenkins R, Hammond K, Spurlock S et al (2022) Separating facts and evaluation: motivation, account, and learnings from a novel approach to evaluating the human impacts of machine learning. AI Soc. https://doi.org/10.1007/s00146-022-01417-y

Kaluža J (2022) Far-reaching effects of the filter bubble, the most notorious metaphor in media studies. AI Soc. https://doi.org/10.1007/s00146-022-01399-x

Kempt H, Heilinger JC, Nagel SK (2022) "I'm afraid I can't let you do that, Doctor": meaningful disagreements with AI in medical contexts. AI Soc. https://doi.org/10.1007/s00146-022-01418-x

Liscio E, van der Meer M, Siebert LC, Jonker CM, Murukannaiah PK (2022) What values should an agent align with? An empirical comparison of general and context-specific values. J Auton Agents Multi-Agent Syst (JAAMAS) 36(1):23. https://doi.org/10.1007/s10458-022-09550-0

Maas J (2022) Machine learning and power relations. AI Soc. https://doi.org/10.1007/s00146-022-01400-7

Munn N, Weijers D (2022) Corporate responsibility for the termination of digital friends. AI Soc. https://doi.org/10.1007/s00146-021-01276-z

Noble SM, Dubljević V (2022) Ethics of AI in organizations. In: Nam CS, Lyons J (eds) Human-centered artificial intelligence. Academic, Cambridge MA, pp 221–240

Novelli C (2022) Legal personhood for the integration of AI systems in the social context: a study hypothesis. AI Soc. https://doi.org/10.1007/s00146-021-01384-w

Omotoyinbo FR (2022) Smart soldiers: towards a more ethical warfare. AI Soc. https://doi.org/10.1007/s00146-022-01385-3

Pflanzer M, Traylor Z, Lyons J, Dubljević V, Nam CS (2022) Ethics of human-AI teaming: principles and perspectives. AI Ethics. https://doi.org/10.1007/s43681-022-00214-z

Singh MP (2022) Consent as a foundation for responsible autonomy. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI). 36(11):12301–12306. https://doi.org/10.1609/aaai.v36i11.21494

Singh AM, Singh MP (2023) Wasabi: a conceptual model for trustworthy artificial intelligence. IEEE Comput 56(2):20–28. https://doi.org/10.1109/MC.2022.3212022

Slota SC, Fleischmann KR, Greenberg S et al (2021) Many hands make many fingers to point: challenges in creating accountable AI. AI Soc. https://doi.org/10.1007/s00146-021-01302-0

Stenseke J (2021) Artificial virtuous agents: from theory to machine implementation. AI Soc. https://doi.org/10.1007/s00146-021-01325-7

Yazdanpanah V, Gerding EH, Stein S et al (2022) Reasoning about responsibility in autonomous systems: challenges and opportunities. AI Soc. https://doi.org/10.1007/s00146-022-01607-8