A sociotechnical system perspective on AI

Kudina, Olya; van de Poel, Ibo

DOI
10.1007/s11023-024-09680-2

Publication date
2024

Document Version
Final published version

Published in
Minds and Machines

Minds and Machines (2024) 34:21
https://doi.org/10.1007/s11023-024-09680-2

A sociotechnical system perspective on AI

Olya Kudina[1] · **Ibo van de Poel**[1]

The aim of this special issue is to provide a sociotechnical system perspective on Artificial Intelligence (AI). Discussions on AI still often focus on the technology itself rather than on the broader systems in which it functions. This is even true when it comes to the social and ethical issues raised by AI. For example, people often talk about fair algorithms or explainable AI, as if fairness and explainability depend solely on the technology and not also on the broader sociotechnical systems in which AI technologies are embedded.

Consider, for example, an AI algorithm that is used by a government agency to find potential cases of fraud with social welfare. Fairness in this case no doubt partly depends on the algorithm itself; for example does it meet certain fairness metrics as they have now been proposed in the literature (e.g., Mehrabi et al., 2019)? However, there are multiple and conflicting fairness metrics, and in order to decide on which of these to focus, one first needs to know more about the broader context: what are the main unfairness that might occur in this context? Which people are most vulnerable and dependent on governmental decisions and should be protected from unfair decisions?

Even when the choice of fairness metrics is based on contextual considerations, it will make the resulting system not necessarily fair. That will also depend, for example, on the behavior of civil servants and politicians, the political climate with respect to fraud, and on internal rules in the relevant governmental agency. Is there room to deviate from the algorithm's advice? Is the emphasis on finding fraud or avoiding unjust accusations? Are citizens offered the possibility to object to a governmental decision or to provide additional evidence if needed? Also legal rules and institutions will affect the functioning and fairness of the resulting sociotechnical system in which the AI algorithm is embedded. For example, are there independent possibilities to appeal against a decision based on the algorithm's advice?

✉ Ibo van de Poel
i.r.vandepoel@tudelft.nl

[1] Delft University of Technology, Delft, Netherlands

Published online: 12 June 2024

A sociotechnical system perspective is important to better understand how AI systems function, what social, political and ethical issues they raise, and how to best address these issues. It also matters for how to design AI; again, the focus should not solely be on the technology and how effective it is in achieving set goals, but also on how it is expected to function in broader sociotechnical systems of which it becomes part. In this contribution, we first set out what a sociotechnical systems perspective is, we then discuss why it is important and in particular what it can add to other approaches; next, we focus on the role of values in and the moral significance of a sociotechnical system approach. Finally, we discuss the contributions that are part of this special issue.

1 What are sociotechnical systems?

A recent EU-U.S. Roadmap for Terminology and Taxonomy for Artificial Intelligence acknowledges as an integral AI attribute it being a sociotechnical system, "constituted by diverse social, political, economic, cultural and technological factors" (European Commission, 2023: 10). But what exactly are sociotechnical systems?

The notion of 'sociotechnical system' has been discussed in a variety of literature including research on technology in the workplace (e.g., Bednar & Welch, 2020; Cooper & Foster, 1971; Mumford, 2006), systems engineering (e.g., Baxter & Sommerville, 2011; Goode, 1957), science and technology studies (STS) (e.g., Bijker et al., 1987; Hughes, 1987; Suchman, 1987), the philosophy of technology (e.g., Kroes et al., 2006; Ottens et al., 2006; Vermaas et al., 2011), engineering ethics (e.g., Borenstein et al., 2019; Noorman, 2014) and more recently also in research on AI and autonomous systems (e.g., Behymer & Flach, 2016; Jones et al., 2013). While there are distinct differences in how the notion of 'sociotechnical system' is understood in these various literatures, a common denominator is the idea that not only technological factors, but also a wide range of social factors is important.

Two ideas are at the core of a sociotechnical system perspective. The first is the idea that technologies are part of larger systems, be it technical or sociotechnical systems. A system might here simply be understood as a number of elements (e.g. technologies, humans, rules) that somehow are interrelated and function together to fulfill a shared goal or objective. Think, for example, of the energy system, the transportation system or the banking system. Such systems may be distinguished at different levels of aggregation. So we might distinguish a world-wide banking system, but also the banking system in a particular country, or the banking system within a specific bank, even the banking system at a specific location of a bank.

The second idea is that among the elements of a sociotechnical system that are crucial for its operation and functioning are human or social elements. Different approaches seem to focus on different types of social elements and are different in how they conceptualize such social elements, and their relation to technical elements. Here we will elaborate a bit on a proposal from the philosophy of technology (Franssen, 2015; Kroes et al., 2006; Ottens et al., 2006; Vermaas et al., 2011).

Kroes et al. (2006) argue that sociotechnical systems have three basic building blocks, namely technologies, human agents and institutions. Institutions are here not

understood as organizations but as social rules that enable and constrain the interactions between human agents (Crawford & Ostrom, 1995). Such institutions can both be formal, like legal rules, but also be informal, e.g. rules that are based on habit or culture. Institutions typically emerge out of past interactions between human agents but they can, to some extent, also be deliberately designed. Creating new sociotechnical systems may also require creating new institutions, think e.g. of user manuals, operation instructions or insurance policies.

AI systems are also sociotechnical systems, but van de Poel (2020) suggests that AI systems as sociotechnical systems have additional (fundamental) building blocks due to the fact that AI is different from traditional technologies, because it is autonomous, interactive, and adaptive. This means that AI can learn from its environment and adapt itself. Of course, traditional sociotechnical systems are also self-learning and adaptive but they are so through human agents. AI systems also have technical elements that allow learning from the environment. Van de Poel (2020) therefore proposes two additional building blocks for AI systems as sociotechnical systems, namely artificial agents and what he calls technical rules. Artificial agents in this conceptualization are non-human non-intentional agents that can autonomously adapt themselves in response to inputs from the environment. Technical rules regulate the interaction between artificial agents (and other technologies), similar to social institutions in traditional sociotechnical systems.

2 The added value of a sociotechnical systems approach

What is the added value of a sociotechnical system approach to AI? First and foremost, we claim, it helps to better understand AI systems, how they function, what social effects they have, what ethical issues they raise, how we can address these issues and how to design better AI. It does so by raising awareness that AI is embedded in broader sociotechnical systems that determine its functioning and effects.

A sociotechnical system perspective is not just important to better understand how AI functions, but also to see how AI might transform existing sociotechnical systems, like the energy system or the transportation system, but also ultimately society and crucial institutions like the law and democracy. If it is true that AI adds new fundamental building blocks to sociotechnical systems as suggested above, it will transform existing sociotechnical systems in a more fundamental way than other technological innovations.

In other words, AI may be a disruptive technology when it comes to other sociotechnical systems and society (Hopster, 2021; Van de Poel et al., 2023). AI may add new elements to energy systems allowing, for example, forms of real-time trading of energy, or optimizing energy delivery through the net. While this may be promising in terms of performance and sustainability, it may also create new vulnerabilities and risks, which requires a rethinking of (the design of) such systems.

AI systems may also disrupt social practices and institutions. Social media, for example, have already disrupted the functioning, but also the traditional understanding, of democracy (Ziliotti et al., 2023). Social robots have challenged the distinction between human and non-human entities like robots (Nyholm et al., 2023), and

ChatGPT is currently disrupting existing practices of writing and education (Shidiq, 2023).

Coming to terms with this disruptive potential of AI is not just a technical challenge. It may require new user practices, new operator skills, changes in existing sociotechnical systems, new institutions and new laws, and perhaps even new moral, legal and philosophical concepts. A sociotechnical systems perspective might be helpful here because by its nature it provides an integral approach to such issues.

More concretely, a sociotechnical systems approach might add three elements that are missing, or are at least less prominent, in other approaches: institutions, culture and governance. Institutions are important because they are consequential for the functioning of AI systems, particularly for how humans interact with each other and with technical elements. However, even some approaches that stress the human element in interactions with AI, like human computer interaction (HCI) and hybrid intelligence do not always pay attention to the role of social institutions (cf. Akata et al., 2020; Norman, 2018).

A second element we want to highlight is culture. Culture may perhaps be seen as an informal institution, but we think it is worth highlighting as it is an element that is not always given due attention in the analysis of AI systems. Cultural constructs, such as beliefs and expectations are inevitably reflected in the training datasets for AI systems, while at the same time the deployment of AI systems may not only reinforce these constructs, but also challenge them.[1] Culture, then, with regard to AI as sociotechnical systems does not presuppose simply a norm-giving environment to the functioning of the system but is an active component to the way an AI system gets appropriated and how effective it is deemed.[2]

A third element is governance. There is now a lot of attention for AI ethics, and rightly so, but properly dealing with the ethical and social issues raised by AI systems is not just a matter of ethical guidelines and proper design, it also requires attention for broader governance and political issues. Given the complexity of sociotechnical systems, dealing with the disruptive potential of AI requires more than just ethics, it requires a range of technical, social, economic and governance choices that are coordinated; in other words, it requires politics.

Finally, a sociotechnical systems perspective has consequences for the design of AI systems. For one thing, it means that AI needs to be developed and designed with an eye to the sociotechnical systems it will be embedded in. For example, how we should understand values like fairness and explainability in the design of AI systems will partly depend on the specific context and the concrete sociotechnical systems it will be embedded in. Designers should not only rely on general design principles or general fairness metrics, but should also tailor these concerns to the specific context and the sociotechnical systems in which the relevant AI technology is to function.

Another design implication is that design should not be confined to the technological part of AI systems, but should also address other elements of the system, like institutions. Some of these elements - like formal institutions - can be designed, but

[1] https://www.theverge.com/2023/2/9/23592647/ai-search-bing-bard-chatgpt-microsoft-google-problems-challenges Accessed 17-10-2023.

[2] https://www.statnews.com/2017/09/05/watson-ibm-cancer/ Accessed 17-10-2023.

oftentimes sociotechnical systems cannot be designed from scratch and sometimes their dynamic might be so complex and emergent, that they can - strictly speaking - not be designed. However, in such cases, often interventions in such systems can still be designed, and the question might shift to how to design such interventions. Anyway, a sociotechnical systems perspective broadens the design perspective on AI substantially.

3 Values and moral significance

Viewing AI through a sociotechnical systems lens is also important because it helps to consider the moral significance of these systems in a broader and more complex sense than when focusing on the technological dimension of AI alone. Just as a purely technical view on AI systems is currently being challenged, so is the conceptual and ontological view on morality as a purely human affair. Instead, a move towards considering morality as a hybrid affair between humans and their broader sociocultural and material environment is gradually becoming accepted in the field of philosophy of technology (Boenink et al., 2010; Vallor, 2016; van de Poel & Kudina, 2022).

On this view, technologies are not neutral tools to be used by people but by virtue of their design features help to give shape to specific moral intuitions, avenues for moral decisions (Verbeek, 2011) and even the infrastructure for moral decision-making - the meaning of the values themselves (Kudina, 2023). Institutions, norms and beliefs, or the broader sociocultural setting of the world is equally not a passive background against which specific human-AI practices play out but helps to give them shape in terms of their nature, process and substance. What follows from the thesis on the intertwined nature of morality with the sociomaterial setting is that even though people remain the moral agents considering morally problematic situations and making decisions regarding AI systems, these same AI systems are not divorced from the moral decision-making of people.

Here a pragmatist take on morality might be especially helpful in underscoring the dynamic interrelation of values and technologies. As van de Poel and Kudina (2022) show, values have a dual nature: on the one hand, they serve as moral orienteers in problematic situation, giving guidance on how to resolve it, and on the other, they are themselves a product of these human-technological practices, which give rise to them, can challenge their fitness or propose new values. As such, values, and by extension the moral problems and moral opportunities are a dynamic property that materializes within a complex system of human-technology-world dimensions and across these dimensions.

With regard to AI, focusing on one of the system components alone would preclude one from having an overview of the larger moral significance of this technology and limit the space for intervention towards a more desirable use or design practice. Equally, any such intervention would need to acknowledge a co-productive way in which the technological component of AI systems presupposes specific moral orientation and a broader worldview as the technological and normative dimensions of AI in the systems view are always intertwined (Gabriel, 2020).

4 Introducing the contributions

The contributions to this Special Issue set out to expand and contextualize our considerations on AI as complex sociotechnical systems. The Special Issue comprises five research papers, each detailing an aspect of AI conceived as sociotechnical systems and presents a broader outlook on the responsible design and integration of such systems.

In their paper "Contestable AI by design: towards a framework," Alfrink and colleagues put contestability as a central feature of the interwoven character of AI systems. Understood as a design feature that underlies the interaction of the social, technological and institutional components of an AI technology, contestability here assumes a system's responsiveness to user's challenge of the system and embeds human rights to self-determination and control in the system. Relying on a mix of methods, such as systematic literature review, the authors propose a framework that would combine essential features (e.g. human review and intervention requests) and practices (e.g. agonistic approaches to ML development) in the AI system to make it contestable. While this framework does not pretend to be exhaustive, it points to the challenges of the decentralized sociotechnical AI systems and at the same time, offers opportunities for system improvement by relying on conflicting points in AI development, leveraging dissent as a rich resource.

In "Spotting when algorithms are wrong," Buijsman and Veluwenkamp tackle the problem of epistemic dependence on AI systems when users have to make critical decisions based on the system's output without being able to verify its credibility, often problematizing the decision-making. Against the background of the Dutch childcare benefits scandal, where a Dutch tax agency employed an AI system that erroneously classified many citizens into high risk categories, the authors propose to tackle the problem of epistemic dependency with introducing system warnings that would reduce people's tendency to trust the AI's output, a mechanism that the authors call "defeaters." Based on the sociotechnical exigencies of AI systems and drawing on the methods to reduce errors in aviation practices, the authors identify five types of such defeaters, ranging from system's inability to provide an accurate output to being based on conflicting statistical correlations. The authors argue that even though an introduction of defeaters into the design of AI systems would require a lot of design consideration and technical challenges, it is worthwhile as it would allow the users to make more nuanced decisions based on these systems.

In "Realising Meaningful Human Control over Automated Driving Systems: a Multidisciplinary Approach," Santoni de Sio and colleagues review the existing work on meaningful human control to provide an operational framework for applying this work to concrete engineering cases. Against the background of a case of Automated Driving Systems, the authors aim to mitigate the responsibility gaps in the use of AI systems related to blameworthiness, moral accountability, public accountability and active responsibility. The two main features in the framework are "tracking," concerning the identification and response to the intentions of the humans operating an AI system, and "tracing," a systems feature that concerns the capacity for and allocation of responsibilities based on the degree of involvement in the decision-making in the AI system. In a transdisciplinary effort, the authors identify the qualifying criteria

to apply this framework such that it is able to systematically locate responsibility gaps and give an idea of the mitigation step, while at the same time providing careful considerations regarding the limitations and further work into the elaboration of meaningful human control.

In their paper "Democratizing AI from a sociotechnical perspective," Noorman and Swierstra propose the lens of political philosopher Mark Warren to develop a democratically legitimate framework of AI development. The authors specifically discuss the cases of AI development and deployment in cases of critical infrastructure, such as in the energy domain, to showcase the democratic stake and complexity in the decision-making between multiple stakeholders. The authors outline specific practices and challenges to democracy that AI, conceived as a sociotechnical system, presents in the energy sector, and propose how these can be mitigated through Warren's approach, focused on three functions of democracy: inclusion, will formation and decision-making. Complementing this lens with the approach to AI as sociotechnical systems building on traditions in Science and Technology Studies and philosophy of technology allows looking at democratization of AI from a broad perspective and accounting for the non-neutral role of AI in the shaping of the democratic domain in this regard.

Finally, in "Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context" Dobbe and Wolters examine AI systems conceptually and empirically to identify their uniquely sociotechnical vulnerabilities and challenges, and to make design suggestions. The authors examine the lifecyle of Machine Learning systems in the financial sector in the experimental, operational and deployment stages, or MLOps, to identify and qualify the AI challenges from the sociotechnical systems standpoint, and propose a set of ten guidelines to mitigate them and to accommodate a broader sociotechnical view on AI in the MLOps practices in the finance industry.

Funding Acknowledgement This paper and the topical collection are a part of the project Value Change, which has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 788321.

References

Akata, Z., Balliet, D., Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., Gaag, L., Harmelen, F. v.,., & Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable Artificial Intelligence. *Computer*, *53*(8), 18–28. https://doi.org/10.1109/MC.2020.2996587.

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, *23*(1), 4–17. https://doi.org/10.1016/j.intcom.2010.07.003.

Bednar, P. M., & Welch, C. (2020). Socio-Technical perspectives on Smart Working: Creating Meaningful and Sustainable systems. *Information Systems Frontiers*, *22*(2), 281–298. https://doi.org/10.1007/s10796-019-09921-1.

Behymer, K. J., & Flach, J. M. (2016). From Autonomous systems to Sociotechnical Systems: Designing Effective collaborations. *She Ji: The Journal of Design Economics and Innovation*, *2*(2), 105–114. https://doi.org/10.1016/j.sheji.2016.09.001.

Bijker, W., Hughes, T. P., & Pinch, T. (Eds.). (1987). *The Social Construction of Technological Systems. New directions in the sociology and history of Technology*. MIT Press.

Boenink, M., Swierstra, T., & Stemerding, D. (2010). Anticipating the interaction between technology and morality: A scenario study of experimenting with humans in bionanotechnology. *Studies in Ethics Law and Technology*, *4*(2), 1–38.

Borenstein, J., Herkert, J. R., & Miller, K. W. (2019). Self-Driving cars and Engineering Ethics: The need for a System Level Analysis. *Science and Engineering Ethics*, *25*(2), 383–398. https://doi.org/10.1007/s11948-017-0006-0.

Cooper, R., & Foster, M. (1971). Sociotechnical systems. *American Psychologist*, *26*(5), 467–474. https://doi.org/10.1037/h0031539.

Crawford, S. E. S., & Ostrom, E. (1995). A grammar of institutions. *American Political Science Review*, *89*(3), 582–600. https://doi.org/10.2307/2082975.

European Commission (2023). *EU-U.S. Terminology and Taxonomy for Artificial Intelligence*. Retrieved 28-10-2023 from https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence.

Franssen, M. (2015). Design for Values and Operator Roles in Sociotechnical Systems. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (pp. 117–149). Springer Netherlands. https://doi.org/10.1007/978-94-007-6970-0_8.

Gabriel, I. (2020). Artificial Intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2.

Goode, H. H. (1957). System engineering: An introduction to the design of large-scale systems. McGraw-Hill Book Co.

Hopster, J. (2021). What are socially disruptive technologies? *Technology in Society*, *67*, 101750. https://doi.org/10.1016/j.techsoc.2021.101750.

Hughes, T. (1987). The evolution of large Technological systems. In W. Bijker, T. Hughes, & T. Pinch (Eds.), *The Social Construction of Technological Systems; new directions in the sociology and history of Technology* (pp. 51–82). MIT Press.

Jones, A. J. I., Artikis, A., & Pitt, J. (2013). The design of intelligent socio-technical systems. *Artificial Intelligence Review*, *39*(1), 5–20. https://doi.org/10.1007/s10462-012-9387-2.

Kroes, P., Franssen, M., Van de Poel, I., & Ottens, M. (2006). Treating socio-technical systems as engineering systems: Some conceptual problems. *Systems Research and Behavioral Science*, *23*(6), 803–814. https://doi.org/10.1002/sres.703.

Kudina, O. (2023). *Moral Hermeneutics and Technology: Making Moral Sense Through Human-Technology-World Relations.* Lexington Books.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. https://doi.org/10.48550/arXiv.1908.09635.

Mumford, E. (2006). The story of socio-technical design: Reflections on its successes, failures and potential. *Information Systems Journal*, *16*(4), 317–342. https://doi.org/10.1111/j.1365-2575.2006.00221.x.

Noorman, M. (2014). Responsibility practices and Unmanned Military technologies. *Science and Engineering Ethics*, *20*(3), 809–826. https://doi.org/10.1007/s11948-013-9484-x.

Norman, K. L. (2018). *The Wiley handbook of human computer interaction* (First edition. ed.). John Wiley & Sons.

Nyholm, S., Friedman, C., Dale, M. T., Puzio, A., Babushkina, D., Löhr, G., Gwagwa, A., Kamphorst, B. A., Perugia, G., & IJsselsteijn, W. (2023). Social robots and society. In I. Van de Poel, L. E. Frank, J. Hermann, J. Hopster, D. Lenzi, S. Nyholm, B. Taebi, & E. Ziliotti (Eds.), *Ethics of socially disruptive technologies* (pp. 53–82). Open Book. https://doi.org/10.11647/obp.0366.03.

Ottens, M., Franssen, M., Kroes, P., & Van de Poel, I. (2006). Modeling engineering systems as socio-technical systems. *International Journal of Critical Infrastructures*, *2*, 133–145.

Shidiq, M. (2023). The Use of Artificial Intelligence-based Chat-GPT and its challenges for the World of Education; from the viewpoint of the development of creative writing skills. *Proceeding of International Conference on Education Society and Humanity*, *1*(1), 353–357. https://ejournal.unuja.ac.id/index.php/icesh/article/view/5614.

Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

van de Poel, I. (2020). Embedding values in Artificial Intelligence (AI) systems. *Minds and Machines*, *30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4.

van de Poel, I., & Kudina, O. (2022). Understanding Technology-Induced Value Change: A pragmatist proposal. *Philosophy & Technology*, *35*(2), 40. https://doi.org/10.1007/s13347-022-00520-8.

Van de Poel, I., Frank, L. E., Hermann, J., Hopster, J., Lenzi, D., Nyholm, S., Taebi, B., & Ziliotti, E. (Eds.). (2023). *Ethics of socially disruptive technologies: An introduction*. Open Book. https://doi.org/10.11647/OBP.0366.

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. The University of Chicago.

Vermaas, P., Kroes, P., van de Poel, I., Franssen, M., & Houkes, W. (2011). *A Philosophy of Technology: From Technical Artefacts to Sociotechnical Systems* (Vol. 6). https://doi.org/10.2200/S00321ED1V01Y201012ETS014.

Ziliotti, E., Benavides, P. D. R., Gwagwa, A., & Dennis, M. J. (2023). Social Media and Democracy. In I. Van de Poel, L. E. Frank, J. Hermann, J. Hopster, D. Lenzi, S. Nyholm, B. Taebi, & E. Ziliotti (Eds.), *Ethics of Socially Disruptive Technologies* (pp. 33–52). Open Book Publishers. https://doi.org/10.11647/obp.0366.03.