Risk Society and AI: Understanding Trustworthy AI at a Societal Level

Asbjørn Følstad
SINTEF
Oslo, Norway
asf@sintef.no

Petter Bae Brandtzaeg
University of Oslo & SINTEF
Oslo, Norway
p.b.brandtzag@media.uio.no

ABSTRACT

Risks of artificial intelligence (AI) may concern the totality of AI as provided by myriad vendors and taken up on a societal scale. In consequence, it becomes increasingly important to address trustworthy AI at a societal level. In this position paper, we discuss such a societal perspective on trustworthy AI. To support this perspective, we present Beck's theory of Risk Society, which concerns how society is increasingly shaped by the identification and management of risks from technological development. We explore how this theory can help understand trustworthy AI at a societal level and detail two key implications. Specifically, we argue that the theory of Risk Society entails (a) the importance of evaluating AI trustworthiness at a societal level and (b) the benefit of open research on trustworthy AI to foster public trust in AI by showing that risks are being actively studied and addressed.

CCS Concepts

• Human-centered computing → Human computer interaction (HCI)

Keywords

Human-centred explainable AI; trustworthy AI; critical infrastructure

1. INTRODUCTION

The rapid advances in artificial intelligence (AI) make it relevant for an increasingly broad range of contexts and tasks. AI is also increasingly taken up for tasks that previously were seen as bottlenecks for computerization, tasks that traditionally have required specific skills and expertise, human judgement, or creativity. Furthermore, AI is taken up in domains critical for societal welfare and value creation, such as healthcare, critical infrastructure, commerce, industry, education and knowledge-work. In consequence, there is a growing importance of trustworthy AI.

Trustworthy AI refers to AI systems developed and validated as providing outcomes desired by users and stakeholders without unforeseen undesirable implications [16]. In line with the aim to avoid unwanted implications, current approaches to trustworthy AI may apply a risk-based approach [18] in acknowledgement of the close relation between AI trustworthiness and risk.

However, while approaches to trustworthy AI typically address the trustworthiness of single AI systems, important AI risks may have implication for the whole society. That is, since AI risk may concern the totality of AI as provided by myriad vendors and taken up by individuals and organizations at societal scale, addressing trustworthy AI mainly from the perspective of single AI systems may limit our awareness and understanding of trustworthiness at a societal level.

To address this limitation, we argue that the theory of Risk Society may be a useful starting point. On basis of this theory, we may construe risks associated with AI as societal without these being due to a clearly defined set of threat actors, paving the way for a societal perspective on trustworthy AI. In the remainder of the paper, we first provide an overview of trustworthy AI and the role of risk management in this context, as well as the need for a societal perspective. Following this, we present the theory of Risk Society for understanding trustworthy AI at a societal level, before detailing two implications of this perspective.

2. TRUSTWORTHY AI

While definitions of trustworthy AI vary, there is broad agreement that trustworthy AI concerns having AI help achieving user and stakeholder goals while mitigating negative or unwanted implications. As summarized by the European Commission High Level Expert Group on AI (HLEG AI), AI trustworthiness concerns AI as lawful, ethical, and robust [7]. Expanding on this, researchers and policymakers have detailed partially overlapping sets of AI trustworthiness requirements or characteristics in need of particular attention, including, e.g., technical robustness, privacy, transparency, fairness, safety, and security [6,7,15,16,18]. To achieve trustworthy AI, significant work has been undertaken on how to set up design and development processes so as to ensure trustworthiness throughout the AI lifecycle [16]. Furthermore, AI trustworthiness may be assessed by verification procedures related to discernible trustworthiness characteristics [24]. Hence, in the words of Kaur et al. [15], trustworthy AI concerns meeting user and stakeholder expectations in a verifiable manner.

Trustworthiness is closely related to risk. Aiming for trustworthiness requires acknowledging relevant risk, and vice versa. Kaur et al. [15] frame trustworthy AI as a means to mitigate AI risk. Liu et al. [17] see avoidance of risk of harm as a key defining characteristic of trustworthy AI. And the HLEG AI [7] states as a key implication of trustworthy AI to maximize benefits while minimizing risks. The coupling of trustworthiness and risk is also made clear in the much-cited NIST [18] AI Risk Management Framework. Here, risk management is construed as key to enable trustworthy AI, addressed through four main functions: Governance, mapping, measurement, and management.

Asbjørn Følstad, Petter Bae Brandtzaeg. 2025. Trustworthy AI in Risk Society. Position paper for the CHI 2025 workshop Sociotechnical AI Governance - STAIG 2025.

From the perspective of risk management, the trustworthiness of single AI systems may be addressed through application of risk management processes and procedures specifically tailored to the domain of AI. The same perspective is reflected in recent European legislation on AI, the AI Act [10], with an objective of promoting the uptake of human-centric and trustworthy AI. Here, providers and deployers of AI systems are obliged to assess their system's risk level and, if found to be within a high-risk category, to establish and maintain a risk management system [ibid., Article 9].

While much research on trustworthy AI concern the development and assessment of single AI systems, a social perspective on trustworthy AI is also represented in the existing literature. This is particularly seen in approaches to trustworthy AI that accentuate sociotechnical aspects of AI [e.g. 19]. As noted by Chatila et al. [6], trustworthy AI is grounded in foundational ethical principles established at a societal level, e.g. in the form of the UN universal declaration of human rights. Furthermore, as for example argued by the HLEG AI in a policy recommendation [8], trustworthy AI requires appropriate governance and regulation. The importance of governance and compliance with policies and regulation is also argued both at the level of individual providers or deployers [18] as well as at a societal level [21]. However, there remains a lack of understanding about how to systematically identify AI risks and assess trustworthiness at a societal level. Moreover, there is a need for clearer guidance on how to translate risk at societal level into concrete policies and regulations. Here, we can learn from the theory of Risk Society.

3. RISK SOCIETY AS A THEORETICAL LENS TO ADDRESS TRUSTWORTHY AI

The theory of Risk Society was proposed by the sociologist Ulrich Beck in the eighties [2]. Beck describes a society increasingly attentive to the problematic aspects of technological progress, drawing on the concept of reflexive modernity. Beck defines risks as systematic and manufactured hazards that emerge as unintended consequences of modernization and technological advancement. While technological progress in the nineteenth and twentieth century often was considered in light of benefits, attention has gradually been drawn towards the immediate problems and potential future risks technological advances may entail. In risk society, the production and distribution of risks is becoming as important as the production and distribution of wealth, and management of risks resulting from scientific and technological advancements has become a key organizing principle.

Characteristic of the man-made risks in risk society is that they are a consequence of technology development uptake rather than the intentions or actions of specific threat actors. Furthermore, they transcend organizational boundaries or national borders, they may impact at all levels of society and they are inherently challenging to control. Due to their diffuse or emerging character, risk following from technological advances and uptake may be challenging to predict. Furthermore, they might lead to popular risk perceptions not to be aligned with actual risk, in some instances leading to overly alarmist risk perceptions whilst in others leading to unwarranted complacency. In a follow-up to Risk Society, World at Risk [3], Beck discusses how common insights into the nature of risk from technological advances could lead to a 'cosmopolitan moment' with opportunities for NGOs and states to align to mitigate existential threats from technology.

Beck does not address digital technology or AI in his discussion of risk society. Rather he address domains such as nuclear energy, industrial pollution, genetic engineering, and climate change. However, we believe that the theory of Risk Society foreshadows key aspects of AI risk, in particular as the current debate on AI concerns swiping societal risks of existential character such as implications on the labour market [9], democracy [23], and humanity at large [5,11].

First, the global and pervasive character of risks in risk society aligns well with those of current AI technology. While leadership in AI development may be held by a few big technology companies, both the prevalence of open approaches to AI, the emergence of advanced applications of AI in ever new small and large companies across the world, as well as the global impact of AI clearly are foreshadowed in theory of Risk Society.

Second, the emergent and diffuse character of AI risk echoes the assumptions of Risk Society. In part as risks associated with AI are located in the future, given further developments towards artificial general intelligence and super intelligence [1], in part as risks and implications have shown difficult to predict, such as the overwhelming impact of large language models on secondary level and higher education [20], while strong claims on the transformational impact of AI on the labour market made right after the public launch of ChatGPT [12] have since then been somewhat moderated [13].

Third, the lack of clearly demarcated responsible actors as well as victims of AI risk corresponds to the ideas of Risk Society. While substantial risk is associated with AI services provided by well-defined actors such as Open AI, Google, Meta, and Anthropic, myriad of other actors provide resembling AI services – several even at competing quality – or AI systems that draw on foundational models provided by others. Likewise, the victims of AI risk may be as broad as to potentially include entire social or demographic groups.

Based on the similarities between AI risks and key assumptions in the theory of Risk Society, we outline two potential implications of this perspective: A need to broaden the scope of trustworthy AI and the promise of open research commons on trustworthy AI and AI risk to foster public trust in AI by showing that risks are being actively studied and addressed.

4. IMPLICATION 1: A NEED TO BROADEN THE SCOPE OF TRUSTWORTHY AI

Applying Risk Society as a theoretical lens on trustworthy AI may be useful to help broaden the scope of what constitutes or reflects trustworthiness for this type of technology. As noted, research and policy on trustworthy AI typically hone in on limited sets of trustworthiness characteristics such as robustness, security, transparency, fairness, and safety [16]. Clearly demarcated sets of trustworthiness requirements are beneficial for development of specific AI systems, as seen from the support provided, for example, in the NIST AI risk management framework [18]. However, such demarcation can also be limiting as important aspects of relevance to AI trustworthiness and risk may not be addressed.

Here, the theory of Risk Society may be valuable as it helps clarify the emergent or diffuse character or AI risk. That is, as AI technology and its uptake evolves, new and unforeseen risks may emerge that are not adequately covered in current approaches to trustworthy AI. This is clearly seen in the developments of large language models, where these – due to their foundational character – may be applied for a broad range of purposes [22] and made use of in applications for a highly diverging domains [4]. Also, in line with theory of Risk Society, there might be divergence between popular risk perceptions and actual risk. For example, while the uptake of AI in higher education by students and teachers has caused substantial concern, recent research also indicate benefits [20].

European guidelines on trustworthy AI may be a good point of departure for such a broadened scope. The HLEG AI [7] accentuates the need to consider trustworthy AI as sociotechnical systems, serves as a basis for policy and legislation, and includes societal and environmental well-being as a key ethical requirement of trustworthy AI. Potentially theory of risk-society may enable further expansion on this requirement, to address this across AI systems and providers.

5. IMPLICATION 2: TOWARDS OPEN RESEARCH COMMONS ON TRUSTWORTHY AI AND AI RISK

Given that AI risk is not limited by organizational or national borders, that it may impact entire groups or demographics, and that there may be no clearly identifiable threat actors, theory of Risk Society may motivate to opening up research and assessments of trustworthy AI and AI risk to involve the community at large. As suggested by Beck, the existential aspect of AI risk may motivate stakeholders of all kinds to join forces to efforts needed to analyze and mitigate risks of AI. Furthermore, given a lack of correspondence between popular perceptions of risk and actual risk, broad involvement and openness may avoid issues of skepticism in the outcomes of assessments run only by researchers and AI experts.

Potentially, establishing open research commons on trustworthy AI and AI risk may be a way forward in part to involve broadly in a challenge going beyond single providers or deployers of AI, and also to enable sufficiently broad consensus on emerging risk to serve as a sound basis for policymaking. By open research commons, we mean organizing the research so as to share, collaborate and involve openly following the pattern of knowledge commons [14]. Opening up research processes on the impact of AI at societal level, by sharing and collaboration on research questions, data, analyses, and knowledge, can help identify risks and advance knowledge on how AI trustworthiness may be achieved at societal level.

6. CONCLUSION

In this position paper we have noted that the field of trustworthy AI may benefit from more closely considering societal implications of AI. We have proposed that Beck's theory of Risk Society may serve as a helpful lens for this purpose. Furthermore, we have shown how trustworthiness and risk associated with AI aligns with key assumptions of this theory, and argue that it can help extend our conceptualization of trustworthy AI. The presented argument is intended as a starting point for further discussions and explorations of how to leverage theory of Risk Society for trustworthy AI.

Through such discussion, we hope to go into detail on how the theoretical lens of Risk Society may best complement other theoretical and practical approaches to AI trustworthiness. Furthermore, it may be relevant to understand whether and how a consideration of the domain of trustworthy AI from the perspective of Risk Society may bring about theoretical advances. As part of this, it will be important to critically consider how insights from application of Risk Society as theoretical lens may impact future method advances and policy development, so as to advance from a promising proposition to a perspective that may substantially advance theory and practice on trustworthy AI.

Acknowledgements: This work is funded by EU Horizon Europe, HORIZON-CL4-2022-HUMAN-02-01, under the project THEMIS 5.0 (grant agreement No. 101121042). The work presented here reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Aschenbrenner, L. (2024). Situational awareness: The decade ahead. https://situational-awareness.ai/

[2] Beck, U. (1992). Risk Society. Towards a New Modernity. SAGE Publications.

[3] Beck, U. (2013). World at Risk. Polity Press.

[4] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[5] Center for AI Safety (2023). Statement on AI Risk. https://www.safe.ai/work/statement-on-ai-risk

[6] Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., & Yeung, K. (2021). Trustworthy ai. Reflections on artificial intelligence for humanity, 13-39.

[7] EC (2019). High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI. European Commission.

[8] EC (2019). High-Level Expert Group on Artificial Intelligence. Policy and investment recommendations for trustworthy AI. European Commission.

[9] Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. Science, 384(6702), 1306-1308.

[10] EU (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689.

[11] Future of Life Institute (2023). Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[12] Goldman Sachs (2023). The potentially large effects of artificial intelligence on economic growth. Economic research.

[13] Goldman Sachs (2024). Gen AI: Too much spend, too little benefit. Global Macro Research, 129. https://www.goldmansachs.com/insights/top-of-mind/gen-ai-too-much-spend-too-little-benefit

[14] Hess, C., & Ostrom, E. (2007). Understanding Knowledge as a Commons: From Theory to Practice. MIT Press.

[15] Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. ACM Computing Surveys (CSUR), 55(2).

[16] Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... & Zhou, B. (2023). Trustworthy AI: From principles to practices. ACM Computing Surveys, 55(9).

[17] 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.

[18] NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, U.S. Department of Commerce.

[19] Polemi, N., Praça, I., Kioskli, K., & Bécue, A. (2024). Challenges and efforts in managing AI trustworthiness risks: a state of knowledge. Frontiers in Big Data, 7, 1381163.

[20] Ravšelj, D., Kerži?, D., Tomaževi?, N., Umek, L., Brezovar, N., A. Iahad, N., ... & Aristovnik, A. (2025). Higher education students' perceptions of ChatGPT: A global study of early reactions. PloS one, 20(2), e0315011.

[21] Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 10(4), 1-31.

[22] Skjuve, M., Brandtzaeg, P. B., & Følstad, A. (2024). Why do people use ChatGPT? Exploring user motivations for generative conversational AI. First Monday, 29(1).

[23] Stockwell, S., Hughes, M., Swatton, P., Zhang, A., Hall, J, & Kieran (2024). AI-Enabled Influence Operations: Safeguarding Future Elections. Research Report. Centre for Emerging Technology and Security, the Alan Turing Institute.

[24] Wing, J. M. (2021). Trustworthy AI. Communications of the ACM, 64(10), 64-71.