This article has been accepted for publication in Public Money & Management, published by Taylor & Francis.

Debate: Evidence-based AI risk assessment for public policy

Thilo Stadelmann[a]*

Outlooks on AI's risks are primarily influenced by proponents' worldviews – what they assume likely or desirable based on prior beliefs, rather than current technological capabilities and development trajectories. For regulating, funding and procuring AI, the public sector needs a more solid foundation of assessing AI than science fiction. This article provides one from the point of view of an engineer and scientist, offering a pragmatic evaluation of ten commonly cited AI risks and respective evidence-based recommendations for policy responses.

Keywords: AI; artificial intelligence; risk assessment; governance; regulation

How worldviews shape AI discourse

AI as a technology and field, with more than 70 years of development history, is reasonably mature (Prince, 2023); the debate surrounding it isn't (Jordan, 2025). Virtually any voice weighing in on risks of and through AI does not argue based on what the technology is or isn't (Stadelmann, 2025), but what they want or fear it to be (Hao, 2025). In part, this is because achieving AI's formulated vision – doing things with a computer that previously could only be done by humans with their minds – included anthropomorphising the technology from the onset, having led to inflated expectations and respective cycles of AI hypes and winters ever since (Segessenmann et al., 2025). But much more, people resort to articulating their *beliefs* on issues where stakes are high, and while this is understandable and even commendable, these views should be subjected to scrutiny before informing public policy. An influential example is the TESCREAL bundle of philosophies (an acronym for transhumanism,

[a]*Centre for Artificial Intelligence, Zurich University of Applied Sciences, Winterthur, Switzerland*

*Correspondence: ZHAW School of Engineering, Thilo Stadelmann, Postfach, CH-8401 Winterthur, Switzerland, stdm@zhaw.ch.

extropianism, singularitarism etc.): Identified as the driver behind much of the loudest voices in the AI field (Gebru & Torres, 2024), it suggests imminent existential risk that drove much of the first governmental reactions to AI but has since been called "ridiculous" by a growing body of leading figures in the field (Heaven, 2023). Meanwhile, tech industry proponents used the ensuing FUD as a strategy that helped their bottom line. What are the implications of such reliance on worldviews for AI governance and public intervention, especially when it comes to assessing AI risks in order to regulate, fund, or procure AI? How could it be done evidence-based instead (van de Poel, 2016)?

A pragmatic, evidence-based approach to AI risk assessment

Many lists of AI risks exist in the literature, with striking similarities; for example, Jackson (2024) lists ten risks of AI. For no particular reason other than its exemplariness, I will use this list to give a brief assessment of each risk with respect to public policy, based on the view that AI is, basically, a normal technology (Narayanan & Kapoor, 2025) and should be treated as such, especially in the absence of any evidence to the contrary (Kambhampati et al., 2025; Kumar et al., 2025).

Existential risk. Alleged existential threats are purely based on the assumption of imminent "artificial general / superintelligence". No plausible scientific path to its creation exists (see references above), and the underlying worldview-based predictions arguably get most of their attention not because of their likelihood, but because of the visibility and economic power of their proponents. Hence, I assess these risks as purely hypothetical; they must not drive current policy.

Dependence. I see our own human dependence on AI systems as a major risk: Giving away freedom of choice and decision-making voluntarily to assumed machine competence as well as sacrificing personal growth in skills and also character through

AI overuse hurt human autonomy. Strengthening moral formation through making people aware of their own value and dignity is a reasonable counter measure (Stadelmann, 2025). This needs to be reflected, for instance, in school curricula and societal institutions: AI is a prime occasion to rethink what (and what for) the human is (Segessenmann et al., 2025), and what we want to keep as inherently human must also guide public funding and procurement. As AI predictions lead to suggestions that ultimately shape human behaviour (de Cárdenas, 2025), dependency's further consequences of, e.g., economic and political nature, finally, need regulatory attention.

A power monopoly. The economics of AI, foremost its business models and ways of distribution, should attract more attention from a regulatory and taxation perspective. As with social media (Orlowski, 2020), the specific way business is conducted in the contemporary tech industry (especially the proclaimed end game of establishing a monopoly and the extreme high stakes in spending) holds much potential for harm and needs to be observed and suitably regulated to avoid similarly bad side effects. Tech sovereignty, i.e., the fundamental freedom of choice of domestic or open source alternatives along the whole AI system value chain, needs to be achieved, which is a task for public funding and procurement.

Ethical dilemmas. Ethical dilemmas are numerous, essentially springing from the conflicts between principles or stakeholders as soon as considering the ethical dimension. For instance, the practice of aligning AI systems with human preferences as an AI safety measure produces "sycophantic" AI systems, making general-purpose chatbots a catastrophic counsellor for serious mental health issues. For the sake of society and the economy, governments need to invest into *prohuman* AI design methodologies and use cases, where the affordances a system offers are morally aligned

with what is good for a human being, to prevent such situations (Schirch, 2023; Acemoğlu, 2025), e.g., as a funding priority and procurement requirement.

Misinformation. It is best to assume that AI generated content is already unidentifiable with any reasonable confidence. I suggest, however, that people adapt socially: Not liking the feeling of being fooled, we will shift our trust back into smaller, more local, more personal relationships. This makes misinformation a lesser public concern apart from criminal offenses.

Job displacement. Many opinions on AI's effect on the labour market are based on the unrealistic assumption of AGI (Narayanan & Kapoor, 2025). Instead, AI does not automate jobs, but tasks, and cannot even come close to automate the "glue" between tasks that makes up most of the complexity in human business. I assume a rather manageable impact on jobs on the societal level but shifts and changes playing out on the individual level. Hence, government can support this in driving curricula in (continuing) education and fostering an evidence-based discourse on the expected change to counter unwarranted fears. Digitally mature citizens are a governmental task, achieved by education. This maturity also helps organizations to make better decisions on what can and should be automated (Ivanova, 2025).

Security threats. AI is a dual use technology that will be used to wilfully do harm. I see an arms race between illegal forces and law enforcement that currently seems reasonably balanced. Governments need to continually support public security institutions to keep pace without drifting into surveillance and authoritarianism. Security threats through not having technology sovereignty have to be addressed (see above).

Data privacy concerns. Existing regulation for example in the EU seems adequate in principle but currently has a hard stand against an industry that largely ignores certain

law and user preferences for the promise of profit (Crofts, 2024). What is and isn't fair use needs to be carefully reconsidered, having not just the most heavily VC-backed stakeholders in mind.

Lack of transparency. Depending on the AI methods used, results are not directly comprehensible for humans (as is also the case with the decisions of fellow humans, and similarly, not every AI result needs an explanation). The field of explainable AI (Bennetot et al., 2024) and new approaches to building machine learning systems (von der Malsburg, 2022; Kumar et al., 2025) offer remedies in the cases where explainability is a reasonable requirement. Regulation should formulate such context-appropriate and stakeholder-adequate requirements, and more R&D effort should be directed into the development of respective methods through public funding.

Bias and discrimination. As far as "algorithmic bias" is concerned, the problem is well understood as rooted in the data (Wehrli, 2022) and finds adequate technical solutions for example in the field and tools of algorithmic fairness (Hertweck, 2023) as well as in existing regulation on non-discrimination. This is no longer an issue of regulation or debate, but of proper implementation (which bears its own challenges).

Recommendations for public policy

According to the assessment above, few risks associated with AI adoption can be considered solved, but all risks appear mitigatable: Short-term risks through technical, social, or by regulatory means as pointed out above and in the literature; longer-term risks by monitoring appropriate evidence. While they appear hypothetical, they should be ignored regarding practice (e.g., policymaking, organizational governance, or the societal debate). The three most prominent recommendations are:

- *Invest* in education on what AI is and isn't; it is the number one government intervention against the major risks of AI dependence and widespread anxiety, with their politically and economically destabilizing effects.
- *Regulate* AI business models with the goal of creating tech sovereignty, also through public *procurement* and open source; these are the most necessary interventions to sustain a fair market and equitable societies where free choice persists.
- *Fund* innovation in the direction of less energy and data-hungry, more common-sensical and fundamentally prohuman AI, which is to be found off the currently beaten path of scaling LLMs; this is the greatest service to higher degrees of transparency, less ethical dilemmas and higher privacy, and should also include incentives for the private sector to invest and build.

Don't fear hypothetical risks; act according to the plenty of evidence instead.

References

Acemoğlu, D. (2025). The World Needs a Pro-Human AI Agenda. *Centre* (Feb. 16, 2025), https://centremag.com/global-affairs/the-world-needs-a-pro-human-ai-agenda/9652/.

Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

de Cárdenas, N. (2025). Franciscan Expert On Artificial Intelligence Addresses Its Ethical Challenges. National Catholic Register (Jan 17, 2025), https://www.ncregister.com/cna/franciscan-expert-on-artificial-intelligence-addresses-its-ethical-challenges.

Crofts, P. (2024). Reconceptualising the crimes of Big Tech. *Griffith Law Review*, *33*(4), 375–399. https://doi.org/10.1080/10383441.2024.2397319.

Gebru, T., & Torres, É. P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*.

Hao, K. (2025). *Empire of AI: Dreams and nightmares in Sam Altman's OpenAI*. Penguin Press.

Heaven, W. D. (2023). How existential risk became the biggest meme in AI. MIT Technology Review (Jun. 19., 2023), https://www.technologyreview.com/2023/06/19/1075140/how-existential-risk-became-biggest-meme-in-ai/.

Hertweck, C., Baumann, J., Loi, M., & Heitz, C. (2023, January). FairnessLab: A Consequence-Sensitive Bias Audit and Mitigation Toolkit. In *EWAF*.

Ivanova, I. (2025). As Klarna flips from AI-first to hiring people again, a new landmark survey reveals most AI projects fail to deliver. Fortune (May 09, 2025), https://fortune.com/2025/05/09/klarna-ai-humans-return-on-investment/.

Jackson, A. (2024). Top 10: Risks of AI. *AI Magazine* (Aug. 21, 2024), https://aimagazine.com/top10/top-10-risks-of-ai.

Jordan, M. I. (2025). A Collectivist, Economic Perspective on AI. *arXiv preprint arXiv:2507.06268*.

Kambhampati, S., Stechly, K., Valmeekam, K., Saldyt, L., Bhambri, S., Palod, V., ... & Biswas, U. (2025). Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces!. *arXiv preprint arXiv:2504.09762*.

Kumar, A., Clune, J., Lehman, J., & Stanley, K. O. (2025). Questioning Representational Optimism in Deep Learning: The Fractured Entangled Representation Hypothesis. *arXiv preprint arXiv:2505.11581*.

von der Malsburg, C., Stadelmann, T., & Grewe, B. F. (2022). A theory of natural intelligence. *arXiv preprint arXiv:2205.00002*.

Narayanan, A., & Kapoor, S. (2025). AI as Normal Technology. 25-09 *Knight First Amend. Inst.* (Apr. 14, 2025), https://knightcolumbia.org/content/ai-as-normal-technology.

Orlowski, J. (2020). The social dilemma. Netflix original (Jul 14., 2025): https://www.thesocialdilemma.com/.

van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and engineering ethics*, 22(3): 667–686.

Prince, S. J. D. (2025). *Understanding deep learning*. MIT press.

Schirch, L., Slachmuijlder, L., Iyer, R. (2023). Toward Prosocial Tech Design Governance. *Tech Policy Press* (Dec. 14, 2023), https://www.techpolicy.press/toward-prosocial-tech-design-governance/.

Segessenmann, J., Stadelmann, T., Davison, A., & Dürr, O. (2023). Assessing deep learning: a work program for the humanities in the age of artificial intelligence. *AI and Ethics*, *5*(1), 1-32.

Stadelmann, T. (2025). How not to fear AI. TEDxZHAW (Jun. 02. 2025), https://youtu.be/deVbP-hViMQ?si=xz39gOt32hVHDg4V.

Wehrli, S., Hertweck, C., Amirian, M., Glüge, S., & Stadelmann, T. (2022). Bias, awareness, and ignorance in deep-learning-based face recognition. *AI and Ethics*, *2*(3), 509-522.

Acknowledgements

I am grateful for constructive feedback by Marcel Blattner and Ricardo Chavarriaga. Regarding GenAI use, DeepL has been used occasionally to find appropriate English phrases and Claude Sonnet 4 has been used on a stable draft of the manuscript to point out weaknesses and suggest titles. As with human feedback, suggestions were analysed and partially inspired improvements; all opinions and remaining errors remain mine.

Author biography

Thilo Stadelmann is Professor of Artificial Intelligence and Machine Learning at the ZHAW School of Engineering in Winterthur/Switzerland, where he is the Founding Director of the Centre for Artificial Intelligence and head of its Machine Perception and Cognition research group. He studied computer science in Giessen and Marburg and received his Doctor of Science degree from Marburg University, Germany, in 2010, where he worked on multimedia analysis and voice recognition. Thilo held engineering and leadership roles in the automotive industry for several years prior to his appointment at the ZHAW, and is is (co-)founder and part of the senior leadership of several organizations in the digital space. His current research interests include representation learning as well as the societal implications of AI and how a hope-filled future for humanity can be attained.