AI and risk: a philosophical analysis

Daniele Chiffi, Viola Schiaffonati, Giacomo Zanotti
Politecnico di Milano, Italy

Abstract

A great deal of attention has recently been devoted to the risks related to the design and use of AI systems. Most notably, the much-discussed European proposal for the *Artificial Intelligence Act* (European Commission, 2021) explicitly builds upon a risk-based approach and provides different levels of regulation for AI systems associated with different levels of risk. Concerns about AI-related risks are motivating a series of initiatives to regulate or pause the development of increasingly powerful AI technologies., such as the recent – and controversial – Future of Life Institute's open letter (FLI, 2023). Yet, discussions on AI-related risks have so far largely ignored the philosophical and scientific literature on risk, and they rarely build upon a proper conceptualization of this notion. This paper aims to fill this gap by providing a conceptual analysis of AI-related risk from a perspective that combines both epistemological and ethical insights.

We start by presenting the main conceptualizations of risk, focusing on the one that is typically employed in the policy literature on natural risk mitigations. Here, risk is conceived in three different dimensions: hazard, exposure and vulnerability (UNISDR, 2015). Then, we proceed by applying this three-dimensional conceptualization to the risks stemming from the use of AI systems. This should allow for a more fine-grained and epistemologically sound analysis of AI-related risk, for different systems arguably involve different risk dimensions. We stress the contraposition between systems whose failure can lead to immediate life-threatening consequences but are used in highly-controlled environments, such as medical AI systems, and systems that involve no immediate threats to life but are nonetheless risky due to their widespread and poorly controlled employment, such as AI-powered chatbots and recommendation systems. Different *kinds* of risk are then distinguished. In addition to loss of lives and material damage, AI systems can involve ethical and social risks – most notably, they can incorporate forms of bias and exacerbate discrimination (see Biddle, 2022). We conclude by arguing that *ex-ante* characterizations of AI-related risks, such as the one at the basis of the proposed AI Act, show severe limitations, for AI development is extremely rapid and often generates contexts of uncertainty (Nordström, 2022). While this element of uncertainty seems to be somehow inbuilt into the development and use of AI systems and can hardly be removed, our fine-grained conceptual analysis enables a better understanding of AI-related risk and thereby more effective mitigation policies.

References

Biddle, J. (2022). On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning. *Canadian Journal of Philosophy*, *52*(3), 321-341. doi:10.1017/can.2020.27

European Commission (2021). *Proposal for a Regulation laying down harmonised rules on artificial intelligence – Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts*. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

Future of Life Institute (2023, Mar. 22). Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Nordström, M. (2022). AI under great uncertainty: implications and decision strategies for public policy. AI & Society, *37*, 1703–1714. https://doi.org/10.1007/s00146-021-01263-4

United Nations Office for Disaster Risk Reduction (UNISDR) (2015). Sendai Framework for

Disaster Risk Reduction 2015-2030. https://www.undrr.org/publication/sendai-framework-disaster-risk-reduction-2015-2030