System Safety and Artificial Intelligence

Roel I.J. Dobbe
Delft University of Technology
Delft, The Netherlands
r.i.j.dobbe@tudelft.nl

ABSTRACT

This article formulates seven lessons for preventing harm in artificial intelligence (AI) systems based on insights from the field of system safety for software-based automation in safety-critical domains. New applications of AI across societal domains and public organizations and infrastructures come with new hazards, which lead to new forms of harm, both grave and pernicious. The text addresses the lack of consensus for diagnosing and eliminating new AI system hazards. For decades, the field of *system safety* has dealt with accidents and harm in safety-critical systems governed by varying degrees of software-based automation and decision-making. This field embraces the core assumption of *systems and control* that AI systems cannot be safeguarded by technical design choices on the model or algorithm alone, instead requiring an end-to-end hazard analysis and design frame that includes the context of use, impacted stakeholders and the formal and informal institutional environment in which the system operates. Safety and other values are then inherently *socio-technical and emergent system properties* that require design and control measures to instantiate these across the technical, social and institutional components of a system. This article honors system safety pioneer Nancy Leveson, by situating her core lessons for today's AI system safety challenges [2]. For every lesson, concrete tools are offered for rethinking and reorganizing the safety management of AI systems, both in design and governance. This history tells us that effective AI safety management requires transdisciplinary approaches and a shared language that allows involvement of all levels of society.

The article is a non-archival contribution to FAccT 2022, and will be published as a chapter to The Oxford Handbook of AI Governance [1]. The full article is available as a pre-print on ArXiv via https://arxiv.org/abs/2202.09292.

CCS CONCEPTS

• **Computer systems organization** → *Embedded and cyber-physical systems*; • **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → *Government technology policy*.

KEYWORDS

artificial intelligence, harms, audits, culture, safety, system safety, governance, policy, automation, systems and control

FAccT '22, June 21–24, 2022, Seoul, South-Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533215

ACM Reference Format:
Roel I.J. Dobbe. 2022. System Safety and Artificial Intelligence. In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3531146.3533215

REFERENCES

[1] Roel Dobbe. 2022. System Safety and Artificial Intelligence. In *The Oxford Handbook of AI Governance*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.001.0001

[2] Nancy G. Leveson. 2012. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, Cambridge, MA, USA. http://ebookcentral.proquest.com/lib/delft/detail.action?docID=3339365