# Designing organizational control mechanisms for consequential AI systems: towards a situated methodology

Shan Amin[1], Roel Dobbe[1] and Sander Renes[1]

[1]Delft University of Technology, Jaffalaan 5, 2628BX Delft, The Netherlands

**Abstract**

Artificial intelligence (AI) holds both potential benefits and significant risks for organizations, including biases, discrimination, opacity, and reduced human accountability. Technical systems, including AI, must be regulated to safeguard stakeholders' interests and maintain proper functioning over time. However, the problem of designing practical controls for specific AI systems and organizations largely remains unresolved. To address this gap, we propose an initial methodology focusing on identifying and contextualizing stakeholders' values within their local environments. We validate our approach through a case study in the Japanese life insurance industry, aiming to assess its repeatability and potential improvements. Our design method includes 10 steps which AI system developers can use to situate high-level institutions in the local context to control their AI systems. The validation efforts highlight the contextual nature of designing controls for AI systems, emphasizing the need for diverse control mechanisms to comply with stakeholders' values.

**Keywords**

Control by design, AI applications, System safety, Design for Values,

Artificial intelligence (AI) offers potential benefits as well as significant risks, including biases, discrimination, opacity, and the reduction of human autonomy and accountability. Therefore, consequential AI systems must be controlled to ensure proper operations over time and eliminate or mitigate known and emerging risks. Recent work on AI safety have emphasized on fixing or amending the AI model or processing its training data or outputs, that is, they have focussed on what you can do prior to putting the system in production. Controlling the system to safeguard values and their associated norms in a production environment, however, requires a more comprehensive approach that extends beyond the model and includes the socio-technical context it is embedded in, as well as the operational and adjacent processes that contribute to its functioning [1]. In the case of AI, in particular AI that has direct impact on human welfare and well-being, additional controls are required to protect stakeholders from harm. AI is a wide concept with many use cases and different stakeholders can have different definitions of harm, therefore, there is no one-size-fits-all solution to the control of AI [2].

Regulatory responses, such as the EU AI act, may provide safety norms or standards, but at this point little guidance exists on how to apply these in particular socio-technical contexts. While a literature has emerged around addressing risks at the level of AI models or at a broader

*Proceedings EGOV-CeDEM-ePart conference, September 1-5, 2024, Ghent University and KU Leuven, Ghent/Leuven, Belgium*

*Corresponding author.

shanamin@outlook.com (S. Amin); r.i.j.dobbe@tudelft.nl (R. Dobbe); S.Renes-1@tudelft.nl (S. Renes)

0000-0003-4633-7023 (R. Dobbe)

governance level, there is currently a gap in literature that aids in conceptualizing and empirically validating the design of control mechanisms for AI systems in their situational context [3, 4].

Our paper addresses this gap by providing a repeatable process for situating identified values into practical organizational controls in the context of specific AI applications. We build our core contribution on earlier approaches that provide ways to tackle parts of the problem. Building on Van De Poel [5], we identify social values and norms relevant for AI systems. Using the safety control structure, a methodology to map and evaluate how different processes relate to the functioning of algorithmic systems [1], we position and operationalize identified norms and values towards requirements on concrete processes and their responsible actors in the form of feedback mechanisms between different processes and their actors. The emerging approach helps identifying a wide range of potential risks, that have to be brought down to a set of feasible and effective requirements for the particular AI application and use case. Building on Garst et al. [6], we propose several steps to reduce the dimensionality of reporting to yield the contextually relevant selection of norms that is controllable for organizations. Furthermore, we then lean on Mäntymäki et al. [7] to further contextualize and incorporate the set of resulting norms within the organization for a particular AI application and use case. This combined approach informs a pragmatic framework for understanding the need and informing the design of organizational control mechanisms for AI systems.

## References


[1] N. G. Leveson, Engineering a safer world: systems thinking applied to safety, Choice Reviews Online 49 (2012) 49–6305. doi:10.5860/choice.49-6305.

[2] R. Dobbe, T. W. Gilbert, Y. Mintz, Hard choices in artificial intelligence, Artificial Intelligence 300 (2021). doi:10.1016/j.artint.2021.103555, article 103555.

[3] M. Anagnostou, O. Karvounidou, C. Katritzidaki, C. Kechagia, K. Melidou, E. Mpeza, I. Konstantinidis, E. Kapantai, C. Berberidis, I. Magnisalis, V. Peristeras, Characteristics and challenges in the industries towards responsible AI: a systematic literature review, Ethics and Information Technology 24 (2022) 37. URL: https://link.springer.com/10.1007/s10676-022-09634-1. doi:10.1007/s10676-022-09634-1.

[4] A. Zuiderwijk, Y.-C. Chen, F. Salem, Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda, Government Information Quarterly 38 (2021) 101577. URL: https://www.sciencedirect.com/science/article/pii/S0740624X21000137. doi:10.1016/j.giq.2021.101577.

[5] I. Van De Poel, Translating values into design requirements, In Philosophy of engineering and technology (pp. (2013) 253–266. doi:10.1007/978-94-007-7762-020.

[6] J. Garst, K. Maas, J. Suijs, Materiality assessment is an art, not a science: Selecting esg topics for sustainability reports, California Management Review 65 (2022) 64–90. doi:10.1177/00081256221120692.

[7] M. Mäntymäki, M. Minkkinen, T. Birkstedt, M. Viljanen, Defining organizational ai governance, AI And Ethics 2 (2022) 603–609. doi:10.1007/s43681-022-00143-x.