Designing Artificial Intelligence for Autonomy

Alfrink, Kars

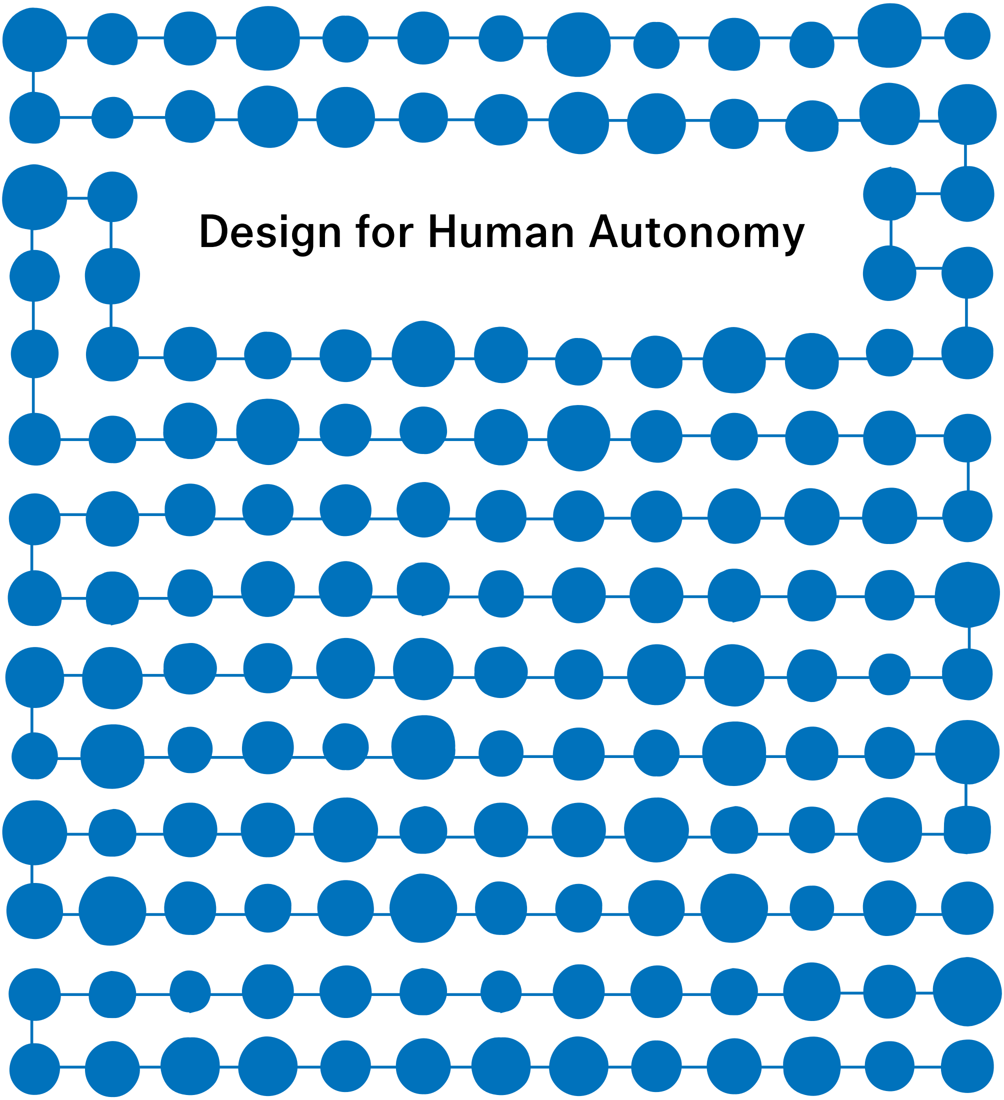
Publication date
2025

Document Version
Final published version

Published in
Design for Human Autonomy

Citation (APA)
Alfrink, K. (2025). Designing Artificial Intelligence for Autonomy. In U. Pesch (Ed.), *Design for Human Autonomy* (pp. 18-19). Delft Design for Values Institute.

Design for Human Autonomy



White paper —— 18 June 2025

Editor: Dr. ir. Udo Pesch

Contributors:

Prof. dr. ir. Ibo van de Poel, Dr. ir. Fatima-Zahra Abou Eddahab-Burke, Dr. Michael Klenk, Dr. Kars Alfrink, Dr. Valentijn Visch, Dr. Victor Muñoz Sanz, Dr. Stephen Rainey, Prof. dr. Tina Comes, Dr. Micah Prendergast, Dr. Simon Parkin

Part of Annual Theme:

Design for Human Autonomy

TU Delft DESIGN FOR VALUES

4.2 Designing Artificial Intelligence for Autonomy

Dr. Kars Alfrink

AI can be understood historically is a subfield of computer and cognitive science. It can also be characterized as a specific set of computational techniques that extract statistical correlations from large datasets, currently dominated by machine learning and neural network approaches. Today, for the most part, these techniques are applied to natural language processing, analysis and generation of 'content' (e.g., text, images, datasets, and programming code), and automated decision/recommendation systems. AI also is a "floating signifier" with strategic vagueness that escapes precise definition while suggesting technological autonomy, serving the interests of its promoters while obscuring the material practices, labor and political economies that make it up. This account is important for our purposes because by treating AI as an "uncontroversial thing" with autonomous agency, rather than a situated set of practices and relations, we contribute to its mystification and shield it from critical examination (Suchman, 2023).

AI poses significant risks to human autonomy, a cornerstone of human dignity. Prunkl (Prunkl, 2022) conceptualizes autonomy as a person's *effective capacity for self-governance*, which in turn depends on two conditions: *authenticity*—holding beliefs free from manipulative influences—and *agency*—acting on these beliefs with meaningful options. AI technologies threaten authenticity through manipulation, adaptive preference formation, and deception. They undermine agency by restricting opportunities, limiting freedoms, diminishing decision-making competence when tasks are outsourced, and imposing paternalistic interventions against individual choice (Prunkl, 2022).

Ethical and rights-based approaches to AI typically address autonomy with principles such as transparency, explainability, accountability, and human control of technology (Fjeld, 2020). However, there are some notable shortcomings in such common approaches to autonomy in AI. First, they tend to emphasize individual rather than collective dimensions. People's effective capacity for self-governance does not depend solely on their person but also on, for example, the institutions and social norms they find themselves enmeshed in. Second, accounts of autonomy in AI (as well as of other values) tend to be universalist in nature, whilst we know that its understanding varies per context. Third, the understanding of autonomy co-evolves with technological developments. The assignment of autonomy to computers by labelling them as 'AI' may, in turn, shape our ideas of what it means for *humans* to be autonomous. Fourth, if it is the case that expectations about autonomy must be negotiated in a situated manner, then a design approach that simply seeks to derive universally applicable interventions from abstract norms will not do. We will have to involve people in the design process. Fifth, and finally, autonomy discourses can be dominated by majority norms—design for autonomy would do well to broaden the scope and be inclusive of a range of 'autonomies.'

Seven preliminary principles can be derived that can guide the design of AI for autonomy, on the level of activities in the design process, and features of the resulting AI system. Activities in the design process should include: (1) reflexively considering how AI technologies reshape our understanding of autonomy itself; (2) examining actual impacts on autonomy of specific material instances of AI systems; (3) creating room for negotiating expectations about autonomy between implicated stakeholders. System features should include: (4) mechanisms for human override; (5) informational resources for transparent operation; (6) non-manipulative interfaces; and (7) support for collective autonomy.

In conclusion, designing AI for autonomy requires a shift from viewing AI as an autonomous entity with agency to recognizing it as a socially situated system embedded within human practices and relations. Rather than imposing universal definitions of autonomy, we must engage in contextual negotiations of what autonomy means across different settings and communities. This approach demands both technical design features that preserve human agency and inclusive design processes that acknowledge the plurality of autonomies. By treating AI as a relational technology rather than an independent actor, we can better ensure that these systems enhance rather than diminish our capacity for authentic self-governance in an increasingly automated world.