Design for Human Autonomy

White paper — 18 June 2025

Editor: Dr. ir. Udo Pesch

Part of Annual Theme:

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

About the authors

Editor: Dr. ir. Udo Pesch
Associate Professor in Ethics and Technology at the Faculty of Technology, Policy and Management, TU Delft.

Prof. dr. ir. Ibo van de Poel
Antoni van Leeuwenhoek Professor in Ethics and Technology at the Facutly of Technology, Policy and Management, TU Delft & Scientific Director of the Delft Design for Values Institute

Dr. ir. Fatima-Zahra Abou Eddahab-Burke
Assistant Professor of Interactive Systems and Engineering Education Design at the Faculty of Technology, Policy and Management, TU Delft

Dr. Michael Klenk
Assistant Professor of Practical Philosophy at the Facutly of Technology, Policy and Management, TU Delft.

Dr. Kars Alfrink
Postdoctoral Researcher at the Facutly of Industrial Design Engineering, TU Delft.

Dr. Valentijn Visch
Associate Professor at the Facutly of Industrial Design Engineering, TU Delft.

Dr. Víctor Muñoz Sanz
Assistant Professor of Urban Design at the Faculty of Architecture and the Built Environment, TU Delft.

Dr. Stephen Rainey
Assistant Professor in Philosophy and Technology at the Facutly of Technology, Policy and Management, TU Delft.

Prof. dr. Tina Comes
Professor of Decision Theory & ICT at the Facutly of Technology, Policy and Management, TU Delft.

Dr. Micah Prendergast
Assistant Professor at the Faculty of Mechanical Engineering, TU Delft

Dr. Simon Parkin
Assistant Professor at the Facutly of Technology, Policy and Management, TU Delft.

White Paper: Design for Human Autonomy

Compiled by the TU Delft Design for Values Institute

This white paper is the result of an interdisciplinary collaboration at TU Delft, bringing together insights from philosophy, design thinking, robotics, and the application of artificial intelligence. It reflects the collective efforts of researchers committed to exploring and advancing the concept of Design for Human Autonomy.

Throughout the 2024–2025 academic year, the Delft Design for Values Institute organized a series of activities centered around this annual theme. The goal was to foster collaboration among TU Delft researchers, promote the practical application of value-sensitive design, and ultimately contribute to a better society.

The white paper is edited by Dr. Udo Pesch, Associate Professor in Ethics and Technology at the Faculty of Technology, Policy and Management, TU Delft. Special thanks are due to Dr. Pesch for his extensive work in connecting disciplines and fostering meaningful dialogue across fields.

We also extend our gratitude to Dr. Michael Klenk, Assistant Professor of Practical Philosophy at the same faculty, who served as the scientific lead for the theme year. Dr. Klenk played a crucial role in bringing together researchers, professionals, and students to engage with the topic of human autonomy in design.

The paper begins by clarifying what is meant by *designing for human autonomy*, followed by an in-depth discussion of the Design for Values methodology. It then explores how this approach is applied across various domains, including technology, policy, and design practice. The concluding section synthesizes insights on different conceptions of human autonomy, the challenges and opportunities in designing for values, and the implications across application domains.

1. Designing for human autonomy

Dr. ir. Udo Pesch

Why design for human autonomy?

With the application of digital technologies, it becomes possible to develop systems that make decisions without human oversight. The question is whether these systems have an impact on the capacity of humans to make their own decisions. In other words, does the autonomy of machines not compromise human autonomy? If so, how can this threat to autonomy be countered by designing human autonomy instead of the autonomy of machines? These considerations have motivated me to single out *design for human autonomy* as this year's theme of the Delft Design Institute for Values.

Human autonomy figures as an ideal for the structure and culture of modern society (Pesch, 2024). The ideal emerged in the Enlightenment of the seventeenth and eighteenth centuries, a period characterized by radical societal, political, and scientific changes. One of these changes has been the transition from a collectivist social order to a social order that prioritizes individual persons (Henrich, 2020). Institutions that used to rely on dogma and hierarchy gradually gave way to institutions like democracy, the legal system, freedom of religion, and the capitalist market. These institutions allowed individuals to shape their own lives based on their personal convictions and preferences.

Still, foregrounding the value of human autonomy might be a big leap at first for designers, as it is not part of the repertoire of values like efficiency, comfort, safety, and sustainability. Autonomy might not be recognized as a value because designers might see it as an empirical characterization of human agency instead of an ethical ideal. Contributing to this idea is the way that autonomy is embedded in modern society; it is so constitutive of our self-understanding and the structure of our world that it has become taken for granted.

Three conceptions of human autonomy

The challenge of autonomous technology brings about the need to design for human autonomy, but while this ideal has been constitutive for the organization of society, it still is an evasive notion. So, before we can think about the design of human autonomy, we need to explore the concept of human autonomy. The starting point of this exploration is that human autonomy is one term, but it captures different meanings that have a systematic overlap, which we will sketch out in terms of three different conceptions of human autonomy.

- The first conception is that of *agency*, in which autonomy is taken as an empirical description of the human capability to make intentional decisions. We have control over our own actions.
- The second conception is that of *freedom*, in which autonomy is related to the political imperative that people should be able to make choices without external interference.
- The third conception is that of *responsibility*, in which autonomy is related to the moral demand that people need to be able to give reasons for why they have made certain choices. This allows the pursuit of individual freedom while maintaining a well-functioning social order.

Taken together, these three conceptions of autonomy inform us that we can be free because we have the ability to make responsible decisions. This idea can be seen as constitutive of modern society, our main institutions are developed so as to facilitate the autonomy of individuals in terms of agency, freedom and responsibility.

The contextuality of autonomy

In our daily lives, we move from one setting to another. We are not just driving cars, but we may be employees, family members, patients, social media users, citizens, students, and customers, and we are constantly negotiating the scope of our agency between different spheres of life. As such, designing for human autonomy implies that we cannot assume a singular conception of autonomy. To address this contextuality of human autonomy, we can follow the work of Joel Anderson (2014), who stated that for each social setting, we can ask ourselves two key questions related to autonomy, namely 'What gets you autonomy' and 'What autonomy gets you'? Anderson gives the example of a driver's license, which is a necessary entrance condition to be able to drive a car; once you have this license, you are free to ride wherever you want to go.

With regard to design, these questions can be articulated as follows: what is the expected range of activities covered within a context, and what is the expected authority and command of people to make decisions within that context? By articulating and discussing these expectations, we can start effectively designing for human autonomy.

Discussing expectations is relevant because otherwise, we may design for the wrong conception of human autonomy. For instance, as Janna van Grunsven and Sabine Roeser (2022) argue, many designs to improve the lives of disabled persons are based on an 'ableist' account of autonomy, which means that technologies are developed that empower people with a disablement to live as much as possible as abled persons – thereby imposing normalcy upon people who may have developed ways of thinking and doing which enable them to pursue their own understanding of autonomy.

In other words, designing for human autonomy implies that it needs to be acknowledged that what is considered to be the scope of agency, freedom, and responsibility will vary per context. There is no fixed meaning of human autonomy, and designers need to account for the heterogeneity and plasticity of autonomy in design for human autonomy. The aim of this white paper is to *explore* how this can take shape in the context of different technologies and design practices.

2. Design for Values

Prof. dr. ir. Ibo van de Poel

Design for Values is a systematic approach for incorporating values of moral importance in the design of new technological artifacts, services and (sociotechnical) systems. It is part of a larger family of value-based design methods including also for example value-sensitive design and value-based engineering.

An important theoretical assumption behind Design for Values approaches is that technology is not value-neutral but value-laden. This means that the social and ethical ramifications of technologies do not only depend on how they are used but also on how they have been designed.

The following phases can be distinguished in a typical Design for Values process: **1)** stakeholder analysis and value elicitation, **2)** conceptualizing and specifying values, **3)** dealing with value conflicts, **4)** prototyping and testing, **5)** ongoing- monitoring and value change. Below these phases are described in more detail.

Stakeholder analysis and value elicitation

The first step in Design for Values is the elicitation of relevant values for which a product, service or system needs to be designed. Sometimes, the focus might be on one specific value but typically, a range of values are relevant for the design of technology. One may look at relevant codes of conduct or legal rules and regulations stipulating values. One might also investigate the expected or potential impacts of a technology, and so define what values should be taken into account in design. Design for Values also requires stakeholders analysis and investigating how stakeholders may be impacted and which values they deem relevant.

Conceptualizing and specifying values

Values are too general and abstract to guide the design process directly. Therefore they need to be conceptualized and specified. Conceptualization of values is “the providing of a definition, analysis or description of a value that clarifies its meaning and often its applicability”. Different conceptualizations of a value may result in different conceptions of it. Specification of values refers to the translation of values into more specific norms and design requirements that can guide the design process. Whereas conceptualization is largely independent from the specific context in which a value is applied, specification is context-specific and aims at specifying want the values means in the specific context in which it is applied.

Identifying and dealing with value conflicts

Although Design for Values may sometimes be focused on one value, it would typically involve a multiplicity of values. In the literature, various strategies for dealing with conflicting values in design have been proposed. When it comes to dealing with conflicts between values, four main types of strategies may be distinguished. The first are *maximizing* or *optimizing* approaches that try to find the best design by making values somehow commensurable. Examples are cost-benefit analysis and multiple criteria analysis. A second possibility is satisficing, which means that a design is sought that is satisfactory, for example by defining threshold requirements that should be minimally met for each relevant value. A third approach is *respecification*. This approach employs the fact that a value can often be specified in a number of different ways. By trying out other adequate value specifications of the relevant values, it might be possible to come to set a set of design requirements that is non-conflicting and still respects the underlying values. A fourth approach might be called *innovation:* it looks at finding new, not yet existing designs that avoid the initial value conflict.

Prototyping and testing

Just like traditional design, Design for Values requires prototyping and testing. With respect to values, prototyping and testing is important to validate whether a proposed design solution indeed embeds the intended values. One way to do this is to test out whether when a prototype is used properly, i.e., according to its use plan, it indeed realizes its intended value. Of course, one might also want to test how robust a design is in realizing or respecting the intended values when used in unintended or unexpected ways. Prototyping and testing might also be used to see whether a design solution has perhaps unintended consequences, which may require taking additional values into account and revising the currently proposed design solution.

Ongoing-monitoring and value change

Oftentimes, the social consequences of - particularly new - technologies only become fully clear when they are already in use. That is to say: despite best efforts to anticipate the social consequences of new technology, and to pro-actively design for values, new technologies may have unexpected consequences that arise during the use phase. To address these properly, on-going monitoring of designed products, services and systems is required. If unexpected and undesirable consequences occur over time, a redesign might be required, or complementary measures might need to be taken.

3. Mapping of human autonomy

Dr. ir. Fatima-Zahra Abou Eddahab-Burke

In the literature on value sensitive design, human autonomy refers to a person's ability to decide, plan and act in ways that support achieving personal goals (Friedman et al., 2013). According to the self-determination theory developed in psychology, human autonomy involves having flexibility and control over own processes and outcomes (Ryan & Deci, 2006). This theory claims that human well-being improves when three universal psychological needs are satisfied: autonomy, relatedness, and competence (Ryan & Deci, 2000). In combination with self-determination, human autonomy is connected to self-rule. It involves beliefs, deliberative processes, and other remaining principles that guide one's action, such as values and desires, as well as choosing between alternatives or agreeing to the interference or guidance by others (Laitinen & Sahlgren, 2021).

Human autonomy, understood as self-determination or as self-rule, is a more demanding concept than pure functional autonomy, which is exemplified by animals or machines like robot vacuum cleaners. Human autonomy refers to the capacity to operate independently without external control and requires a sufficient degree of control over instincts and impulses, which most animals lack (Laitinen & Sahlgren, 2021). This form of personal autonomy is essential to the notion of moral responsibility that represents the ability to assess reasons for action, to pursue values, and to resist irrational impulses (Ripstein, 1999). This goes in the same direction as what was initially claimed by Kant, for whom true autonomy means using reason to guide one's actions, making choices based on moral principles derived from rational thinking, not merely reacting to circumstances or following feelings and impulses (Kant, 1785). Therefore, human autonomy, according to Kant, is opposite to being offhand, unadvised, spontaneous, random, haphazard, casual, or without any regard to the merit of man. Human autonomy requires deliberate thought, not accidental or impulse-driven behavior (Korgaard, 1996; Hill, 2000).

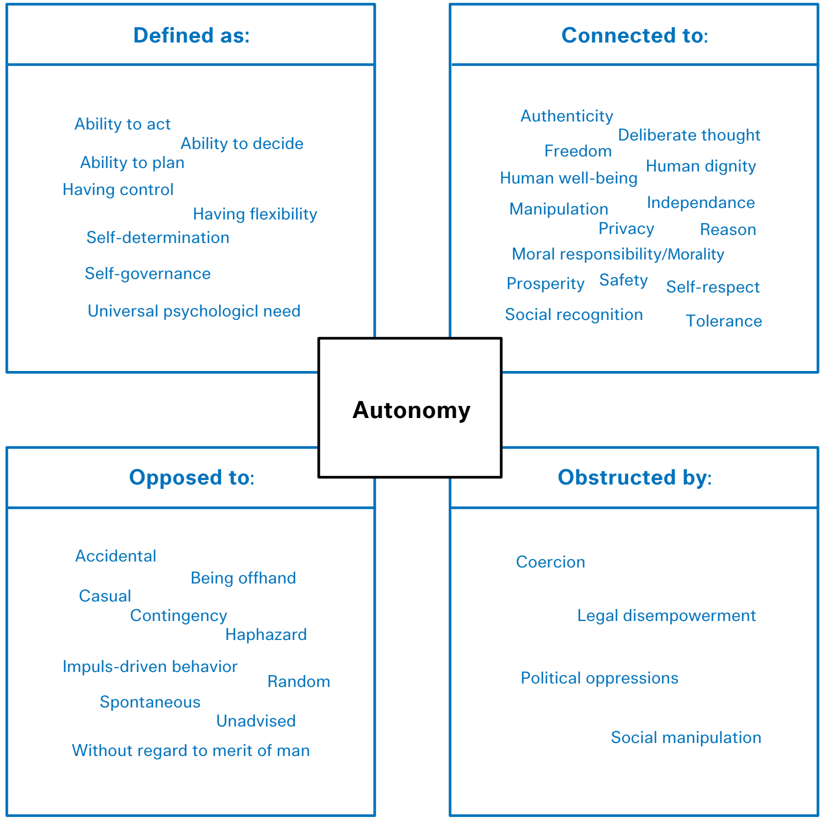
Human autonomy is also found in the literature as self-governance. It refers to the capacity to form one's own understanding of what is valuable and meaningful, to develop the values that will direct one's actions and choices, and to make significant life decisions based on those values where one sees fit (Rubel et al., 2020). This connects human autonomy with independence, defined as the ability to make one's own life choices in accordance with one's objectives and values (Kamphorst, 2012). Another aspect closely related to human autonomy is social recognition. Researchers claim that the social aspect is not just something that helps autonomy develop or be used, it is actually a key part of autonomy itself (Mackenzie & Stojar, 2000) Recognition is important for both building personal potential and healthy self-relationships, as well as for effectively exercising autonomy (Kauppinen, 2011).

We live in a connected world, where interaction with intelligent machines both online and offline has increased significantly over the past decades (Floridi, 2014; Li et al., 2022). This raised several debates about the impact of technology on human autonomy. On the one hand, some researchers claim that technology is a threat to autonomy and can undermine it (Aylsworth & Castro, 2024), and that human interactions with technologies are worrisome as they appear to be highly manipulative that can seriously affect human well-being and autonomy (Susser et al., 2019; Frischmann & Selinger 2019). The reasoning behind this is that being manipulated can make humans act towards ends they haven't chosen or lead them to act for reasons that are not authentically theirs (Susser el., 2019). Other researchers disagree that this is a transition where technologies have gained agency or became potential threats to human autonomy (de Mul & van den Berg, 2011). As mentioned in (Klenk & Hancock, 2019), "an individual can be manipulated while being fully personally autonomous" and technology is capable of manipulating humans without compromising their autonomy. It is acceptable for a technology or a system to take over control as long as it does not interfere with or diminish one's autonomy (Kamphorst, 2012). In some ethical guidelines, autonomy is also called agency to ensure that people have full control over their decisions in interaction with technologies (Agarwal et al., 2022). Also, Higher-Level Expert Group on AI (2019) has formulated four ethical principles that should be met by AI applications, namely respect for human autonomy, prevention of harm, fairness and explicability.

Although autonomy is a complex value (see Figure 1), it should be respected (Raz, 2001), and this principle extends to technologies and systems, which should be both designed and function in ways that preserve human autonomy (Kamphorst, 2012). Autonomy should not only be respected because it is valuable in itself, but also as a constituent of human dignity (Riley & Bos,

2021) and well-being (Ryan & Deci, 2017). However, it can still be threatened by obstacles such as various forms of social manipulation, coercion, legal disempowerment, or political oppression (Laitinen & Sahlgren, 2021). To avoid this situation and do justice to an important value such as "respect of human autonomy" in a technological and digital era, human autonomy and agency should be distinguished from the potential autonomy and agency of AI (van de Poel, 2020).

Figure 1: Mapping of the value of human autonomy



4.1 Measuring Manipulation Workshop Design for Human Autonomy

Dr. Michael Klenk

One of the key issues regarding human autonomy evolves from the development of technologies that enable large-scale manipulation, like social media and generative AI enabling potentially harmful, large-scale manipulation. With that, the topic of manipulative influence has gained prominence in public and academic discourse. In this, philosophical studies offer conceptual insights into manipulation but lack empirical grounding, while scientific approaches measure specific influences, such as misinformation, but seldom engage with philosophical frameworks. In a workshop, researchers from diverse disciplines—including philosophy, psychology, computer science, HCI, and law— were brought together to explore methods for measuring (online) manipulation. The focus of the workshop was on online influences, such as YouTube's recommendation algorithms, while acknowledging that manipulation extends beyond online contexts.

Several researchers presented diverse perspectives on manipulation in human-AI interaction, highlighting both technical and ethical dimensions. One researcher examined the evolution of "direct manipulation" in human-computer interaction, originally designed to empower users through immediate and reversible actions but now contrasted with emerging forms of ethically questionable manipulation. A literature review identified four negatively valanced themes that could serve as a foundation for a broader analysis. Another study explored how conversational agents that suggest arguments can shift users' attitudes, depending on whether the arguments support or oppose a topic like social media. This raised ethical concerns about the intent behind such influence. Filtering and recommendation systems were also discussed, particularly their potential to manipulate vulnerable users such as children, despite their utility in managing information overload. An agent-based model of a book recommender system showed how hidden influence could lead users to adapt their preferences in ways that diverge from their original intentions, potentially causing harm. Research on social behavior in flies revealed mechanisms of learned social flexibility, prompting questions about whether such models can predict manipulation or only measure its effects. An experimental framework was introduced to assess manipulation by comparing participants' stated privacy attitudes with their actual choices in either neutral or manipulative environments, focusing on perceived control. Finally, a philosophical perspective challenged common definitions of manipulation, suggesting that the key factor might be the influencer's indifference to whether their influence is based on valid reasons. This idea sparked discussion on how such a criterion could be tested in practice.

From this discussion, several important questions for future research emerged. Could we create a glossary of manipulation, similar to the *Lexicon of Lies* for misinformation? Do intentions matter in measuring manipulation, and if so, how can they be reliably assessed? Can manipulation be measured in real time rather than retrospectively, perhaps by using predictive models to detect behavioral deviations? What are reliable indicators of manipulation—such as regret, divergence between stated and actual preferences, or outcomes in controlled scenarios? How does context—platform, content, and user characteristics—influence manipulation, and what role does friction (e.g., click-signals versus user-stated preferences) play? Finally, could users' perceived sense of being manipulated serve as a measurable and meaningful indicator?

On Dec 12 – 13 at TU Delft, a workshop was organized by Michael Klenk with the Delft Design for Values Institute and support from AlumNode to examine manipulative influence and address major conceptual and empirical gaps — especially in defining, operationalizing, and measuring manipulation. The reliable detection and measurement of manipulation is especially urgent, as EU regulations like the AI Act explicitly prohibit manipulative influence. The participants of the workshop were:

Siddharth Mehrorta — (TU Delft & UvA) – Postdoc Researcher on trust modelling in human-AI Interaction

Maurice Jakesch — (Bauhaus University Weimar) Assistant Professor Chair of Computational Social Science

Cynthia Liem — (TU Delft) Associate Professor on Trustworthy and Responsible Artificial Intelligence

Javier Mondragón Briseño — (Software Engineer at Freeday) Alumnus of Master Complex Systems Engineering and Management at TU Delft

Frederic Alexander Römschied — (European Neuroscience Institute Göttingen (ENI-G)) Research Group Leader

Friederike Stock — (Max Planck Institute for Human Development) Predoctoral Fellow on Boosting Decision Making

Michael Klenk — (TU Delft) Assistant Professor on Ethics and Philosophy of Influence and Manipulation.

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.

4.3 Autonomy in Practice: A Human-Centered Design Approach to Health Influence

Dr. Valentijn Visch, Dr. Hakan Kuru & Dr. David de Buisonjé

Health is inherently personal. We can even argue that everyone "owns" their health and, as (partly) autonomous beings, they possess agency, freedom, and responsibility over their "own" health. This notion resonates with the contemporary definition of health as a "dynamic and valued level of resilience"—the capacity to adapt and self-manage in the face of social, physical, and emotional challenges (van Vliet et al., 2024). However, the scope of this autonomy is not absolute. Life events such as chronic disease or cancer resulting from living in an uncontrollably unhealthy environment can limit the degree to which people have autonomy over their health.

Even so, there is value in enabling individuals to understand when, how, and to what extent they can exert control. Adopting a healthy lifestyle increases the likelihood of positive health outcomes, but this influence has its limits. Then, autonomy in health is less about complete control and more about meaningful influence within the "own" context. We must also recognize that not everyone has equal opportunity to act on their health, and expecting self-management can reinforce inequality, particularly among those facing limited time, insufficient resources (financial, social, therapeutic), or chronic illness.

In AI-enriched contexts, human-centered design offers a pathway to support autonomous, context-sensitive health management. We propose that such a design can enhance autonomy by focusing on two key areas:

a) Agency-increasing awareness of one's ability to influence health, and

b) Responsibility and freedom-facilitating supporting actions to promote that influence

Designing for Awareness

One way to foster agency is the use of "own" health data. Wearable devices like FitBit, WHOOP or Apple Watch generate biodata feedback—such as heart rate variability or sleep quality—that can be translated into a subjective health indicator (e.g., Fitbit's readiness score or Apple's Vitals Dashboard). These interpretations can give users a sense of ownership and insight, especially when applied with adaptive AI that can contextualize the feedback. AI further enhances this awareness by identifying patterns invisible to the human eye, such as subtle signs of overtraining, early fatigue, or poor recovery. It promotes a sense of informed agency, allowing individuals to notice, reflect, and respond before problems escalate.

However, data alone does not guarantee empowerment. The current challenge lies in making data actionable, comprehensible, and personally relevant. Here, design plays a critical role: visualizations, feedback loops, and AI-guided narratives must be aligned with the user's health goals, values, and experiences. As a result, this approach will inform the user and support autonomous decision-making rooted in informed agency.

Designing for Motivation

Autonomy is not only about knowing; it is also about doing. Effective health motivation design Autonomy is not only about knowing; it is also about doing. Effective health motivation design must go beyond short-term extrinsic rewards (like money) and incorporate intrinsic motivators, such as the fundamental human need for autonomy (cf. Ryan & Deci, 2004).

At this point, several design strategies can enable behavioral freedom and responsibility. One is the concept of **design for crafting** (cf. Wei Liu, in review; Bakker & Oerlemans, 2019), where individuals learn to shape and optimize their environments to support their health-related goals. For example, Apple Fitness allows users to personalize routines and set goals that align with their values, encouraging them to craft a health journey that feels truly their own. This links closely with design for **self-experimentation**, where design encourages individuals to try different approaches and discover the most effective, context-specific ways to influence their health. Wearable devices exemplify this by enabling users to track variables like sleep, strain, or daily routines—allowing them to observe how behavior changes (e.g., avoiding caffeine before bed or shifting workout intensity) impact recovery and energy levels. Over time, users build personal insight through iterative learning. Another one is **contextual adaptation**, which provides tailoring interventions to social and physical environments to ensure that autonomy is both meaningful and feasible. Devices like the Garmin smartwatch adapt training suggestions based on heat, altitude, and past performance, while apps like Woebot adjust emotional support content based on time of day or user sentiment.

In summary, designing for health influence means designing for autonomy, grounded in agency, freedom, and responsibility. By making individuals more aware of their potential impact on health and supporting them with tools and environments that respect their context, health design can move toward empowerment without pressure and support without overreach.

4.4 Human Autonomy, Spatial Design and the Future of Work

Dr. Víctor Muñoz Sanz

True gamers use RGB LED lighting to boost their concentration. “This is what my room looks like with no RGB, I literally feel dead inside,” one explains (Tech Tesseract 2022). This observation might appear trivial, but it illustrates a powerful dynamic: environments—lighting, climate, furniture, and display systems—can significantly shape cognitive and emotional states.

These ideas of stimulation, choice, and play bleed into the ways contemporary work is organized. In automated greenhouses, for example, wearables replace inefficient systems of data registration, while performance rankings prompt competition among workers. The gamification of labor is not merely a metaphor—it is a design strategy.

This brings us to a foundational question in labor studies: “Why do workers routinely consent to their own exploitation?” Building on Gramsci’s theory of hegemony and ethnographic research, Burawoy (1979) argued that management secures control by giving labor the “illusion of choice” within a highly restrictive environment.

In parallel, we must also reconsider how technologies themselves shape this landscape. Digital systems in workplaces are often mislabeled as ‘smart’, ‘automated’, or ‘artificially intelligent’. As such, while many jobs are transforming, they are not being fully replaced (OECD 2016). Instead, they are being reorganized into more complex human-machine entanglements (Autor 2015; Sadowski 2022; Taylor 2018). Humans haven’t disappeared, but rather have been redefined—as “conscious linkages” in an automatic system of “mechanical and intellectual organs” (Marx 1973). Managerial strategies thus aim to secure workers’ cooperation in a system that increasingly limits actual autonomy (Muñoz Sanz 2018).

The changing structure of work continuously reshapes workers' subjectivities. In domains like indoor horticulture and logistics—marked by automation and precarious employment—direct discipline is less effective. Instead, digital interfaces, platforms, and sensors manufacture consent through play, personalization, and the promise of autonomy (Perrig 2021). Under the guise of productivity, these technologies mask manipulative mechanisms that increase workloads and erode autonomy (ILO 2022).

To understand this more deeply, Han's theory of psychopolitics is highly instructive (2017). Smart technologies, in his view, are seductive instruments: they operate not by force, but by emotionally rewarding users. They don't discipline; they gratify. This shift—from coercion to self-regulation—exploits the psychic realm, making individuals complicit in their own control. Such transformations in work and subjectivity are receiving increasing attention in organizational studies, philosophy, and digital geography (Jackson et al. 2006; Mollick & Rothbard 2014; Taskin et al. 2019). Importantly, this research shows that design choices—especially aesthetic and spatial ones—can function psychopolitically. They elicit emotional responses, shape behavior, and induce a disposition of compliance (Lyn et al. 2022; Lynch 2021). In this sense, (spatial) design—and by extension, architecture—becomes a tool of governance in automated landscapes.

This raises a critical question: how might spatial design participate in the psychopolitics of control and worker consent? We need deeper knowledge on how the architecture of automated landscapes transforms subjectivities undermining human autonomy, particularly through psychopolitical tactics. Without examining the shift from design for to design of the user-subject position (Bratton 2015), we risk facilitating the quiet displacement of the human in subtly manipulative environments.

4.5 Neuroenchancement and design for human autonomy

Dr. Stephen Rainey

Neurotechnologies include systems that interface between the human brain, or wider nervous system, and devices or software (OECD, 2020). These interfaces are typically based on recording of neural activity, like the electrical activity of brain cells. Patterns of electrical activity recorded from specific areas can be used to control devices or software (*Neurotechnologies*, n.d.). For example, electrical activity recorded from the brain's articulatory motor system can be used to control software to produce synthetic speech, allowing those with sever aphasia or locked-in syndrome to communicate (Maslen & Rainey, 2020). Prosthetic limbs can be controlled by signals recorded from muscle activity in limbs to replace mobility (Agashe et al., 2016). Brain activity can be monitored by devices that can detect and foreshorten neurological problems like epileptic seizures, through stimulation like a pacemaker (Kavehei et al., 2019). Mood disorders and psychiatric issues can potentially be addressed through administering electrical stimulation deep within the brain (Williams & Okun, 2013). With these few examples, alongside many other use cases, neurotechnology can contribute to the capacity of humans to act autonomously through replacing lost physical abilities, intervening on compromised abilities, or augmenting abilities.

Owing to the complexity of brain signals themselves, not to mention the sheer volume, artificial intelligence (AI) will grow in prominence for decoding brain signal as neurotechnology continues to develop (Badrulhisham et al., 2024). While the ultimate effects of the technology appear to boost autonomy, the inclusion of AI in the workflow opens questions about whether and how users can be said to explain their reasons. In speech, for instance, the inclusion of predictive language models between the recording of brain signal and synthetic verbal output aids in speeding the production of speech. However, a potential gap between the prediction of the model and the intentions of the speaker opens (Rainey, 2023). How much of a gap should be considered acceptable is an open question. Besides AI, deep brain stimulation (DBS) has been reported to coincide with personality changes in users treated for Parkinson's Disease (Klaming & Haselager, 2013). If the explanation for actions and reasons became the DBS intervention itself, that may signal an issue for the autonomy-preserving credentials of DBS approaches. In this case, the autonomy boost of having tremors addressed and mobility and physical activity enabled is offset with the potential for erratic behaviour and obscure motivations.

These two examples symbolise emerging questions, derived from the action of some specific neurotechnologies themselves, about how reason, personality, and motivation interact where autonomy is at stake. In light of reliance upon neurotechnologies for speech or action, and with developing literature on how these operate, a person's expectation for what autonomy is may become fluid. Trade-offs between control, accuracy of intention-translation, and consistency may be expected, and may alter the ways neurotechnology user can be judged in terms of autonomy (Friedrich et al., 2018; Gilbert, 2015; Walker & Mackenzie, 2020).

Neurotechnology design and evaluation may need to prepare for a future in which meanings of autonomy and neurotechnological design shift in tandem. Similarly, philosophers and ethicists may need to adopt pragmatic stances toward neurotechnology users, the reasoning their devices offers them, and the actions that are enabled thereby.

4.6 When Drones Deliver Aid: Human Autonomy in Algorithmic Humanitarian Logistics

Prof. dr. Tina Comes

Humanitarian Drones

Humanitarian responders need to provide assistance to those in need while facing uncertainty, time pressure, high stakes, and severe access constraints (Comes, 2024). Unmanned Aerial Vehicles (UAVS) promise to address several critical humanitarian challenges simultaneously. They offer rapid deployment capabilities that bypass destroyed infrastructure, provide real-time situational awareness through aerial reconnaissance, and enable "last mile" delivery to remote populations. Even more appealing is the promise of UAVs to solve the fundamental tension between speed and safety—delivering aid quickly without exposing humanitarian workers to danger.

Since their first deployment in the Democratic Republic of Congo in 2006, UAVs have rapidly transformed humanitarian logistics, evolving from surveillance tools to autonomous cargo drones delivering vaccines, blood supplies, and HIV diagnostic kits to remote populations (van Wynsberghe & Comes, 2020). Today, the humanitarian drone is set to deliver three capabilities: *transportation* especially solving "*last mile*" access problems, *monitoring* and remote sensing, and *communication* replacing or complementing destroyed terrestrial networks (Rejeb et al., 2021). As such, UAVs fundamentally reshape how humanitarian actors access information, make sense of crises and coordinate responses.

What makes UAVs particularly interesting in the context of autonomy is their dual role for both mediating access to information and as logistics actors: UAVs are expected to at the same time assess damage and needs, and deliver assistance. This convergence of sensing and acting positions UAVs uniquely in the humanitarian AI and robotics space, since they function as agents that fundamentally change information collection, sharing and decision-making authority in crisis response.

Design Principles for Preserving Human Autonomy

UAV systems should provide information to support human sensemaking and decision-making. Following principles of human-centred AI (Shneiderman, 2020) technologies should enhance rather than replace humans. In the context of humanitarian response, this explicitly includes the need for sensemaking, human contact, contextual interpretation and ethical reasoning. This requires ensuring that humanitarian responders retain ultimate authority, responsibility over and accountability for decisions that involve moral trade-offs or community preferences.

UAV integration should incorporate community voices in decisions about when, how, and where autonomous systems are deployed. This is especially true as crises often hit the most vulnerable people hardest. As the immediate need for assistance can open the door for humanitarian experimentation (Sandvik et al., 2017), a priori design and development is needed that clearly state the conditions how and under which circumstances UAVs can be used, how the information will be processed, shared and retained, and how the algorithms that prioritise assistance will work, as well as what are adequate mechanisms for feedback and oversight.

The promise of UAVs to solve humanitarian access challenges must be balanced against their potential to undermine the human autonomy that enables effective, dignified, and context-aware aid delivery. As van Wynsberghe and Comes (2019) conclude, the moral acceptability and desirability of drones in humanitarian contexts is dependent on understanding their full impact on human relationships and decision-making processes. Rather than viewing UAVs as neutral efficiency tools, humanitarian organizations must recognize them as agents that actively reshape social relationships, information flows, and moral reasoning processes and thus deeply impact human autonomy.

4.7 Robotics and Human Autonomy

Dr. Micah Prendergast

Robots with the ability to interact both socially and physically in the world offer limitless possibilities for both enhancing and impeding human autonomy. Design choices that seek the former must consider not only the technical challenges, the needs of human stakeholders and the operational/social context, but also the inherent mechanism by which they hope to enhance autonomy. Formosa et al. (2021) provide a useful framework for considering how social robots (and in quite some ways physically assistive and collaborative robots as well) may enhance human autonomy. They argue that robots may enhance human autonomy via at least three distinct mechanisms: (1) by providing either a means to achieve some objective directly via the robot's actions or the means to achieve a more valuable objective, (2) by improving autonomy competencies of the human and (3) by allowing the human to make more authentic choices.

The first of these mechanisms may include physically assistive robots designed to enable those with restricted functions either by performing tasks for them (Kemp et al., 2022) or providing them restored function in some way. This could include robots that assist mobility-limited humans within a home (van Nes et al., 2022; Song et al., 2024; Farina et al., 2023), physically assist humans within the workplace, and robots that facilitate more valuable human objectives by expediting or performing less valuable ones (Tribelhorn & Dodds, 2007). Augmented human function would also fall within this category, providing new more valuable objectives that would have otherwise been impossible for the human to perform. These include robotic lifting aids that provide superhuman strength (Kirkwood et al., 2022), surgical robots with precise control that exceed the physical abilities of humans (Chatterjee et al., 2024), and physically interactive robots that may perceive and respond to human physiology in a way that is beyond human perception and control (Prendergast et al., 2021).

The second of these autonomy enhancing mechanisms may be best thought of as a robot providing the opportunity for a human to further their autonomy. This is achieved by handing off mundane or autonomy limiting tasks to a robot such that the human may improve their own knowledge/skills and thus their ability to be autonomous. This is distinct from the first mechanism in that the human is not simply trading off one less valuable objective for a more valuable objective, but rather a less valuable objective for the objective of directly enhancing their autonomy competency. A robot that does the cleaning (Kemp et al., 2022) or mows the lawn (Wu et al., 2022) so that a human has time to improve or learn a new skill, be it music, welding, dancing etc., serves as a nice example.

The third mechanism introduces the concept of authenticity with the goal of enabling more authentic choices for the human. Socially this could take the form of a robot helping a human make choices and decisions that align with their values. Physically assistive robots may also enable autonomy in this way, providing a means for humans to express themselves physically and to better enact their physical preferences whether it be to overcome a loss of function, or to gain physical ability (Huynh et al., 2021).

The diversity of robotic capabilities provides the potential for enhancing human autonomy in powerful ways. Unfortunately, that same power also provides a means for the manipulation and exploitation of humans and for directly inhibiting human autonomy. Robots that seek to enhance human autonomy must be designed with a consideration for a high dimensional landscape of social contexts, human stakeholders and relationship dynamics in addition to simply functioning as desired. In addition, while seeking to design robots for enhancing autonomy we run the risk of creating new and unexpected tasks for ourselves that may serve to reinforce existing social biases and inequality (Coggins, 2022). Transdisciplinary efforts that directly engage the many relevant disciplines and stakeholders as a part of these development processes offer a strong foundation for ensuring that robotic technology enhances rather than impedes human autonomy (Zaga et al., 2024).

4.8 Online security and human autonomy

Dr. Simon Parkin

Automated decisions favor the security of the overall system, rather than allowing a human to 'decide their own fate' and take their own risks (in essence, security of the system, over securing what the user wants to do). We see this playing out in a mundane way every day, when legitimate emails are silently hidden away in Junk folders. We also see it when browsers enact the 'opinionated design' of the browser provider, warning users not to proceed to a website which 'might' be malicious, and at the same time do not suggest a way for the user to reach their original goal (for instance, buying concert tickets) – 'success' is deemed to be simply stopping the user from taking a risk, regardless of whether they achieved their goal or not in the end. And yet, expert practitioners are not able to agree on the key steps users should take to stay safe when navigating online spaces. The point here is that remaining secure is seen as being qualitatively equivalent to any other outcome a user would have wanted. This is the same process which can habituate users to missing out on the benefits of technologies, such as avoiding online banking for fear of the risks.

A lack of humans being able to do what they would have done is not noticed, and is of potentially little interest to the 'owner of' the security of an ecosystem, because a user doing nothing and forgoing an activity which exposes them to a risk - rather than doing something with a cyber-risk associated with it - preserves the confidentiality and integrity of a system, while also seeming to maintain availability (even if it is not realized). Security is a secondary task, and without consideration of the user's primary goal, a dilemma may be imposed upon a user, where they have to decide between conflicting outcomes. There is, however, potential to design systems so that productivity and security goals can be met together.

This relates to a concern about *unintended harms* in cyber-risk management. Those practitioners responsible for managing cyber-risk will have good intentions – namely, to secure the system – but having good intentions does not insulate our actions from having potential negative consequences. Unintended harms in cyber-risk management can include displacement of legitimate users from a service, burdening of users with additional effort, amplification of existing harms, etc. This can extend to those instances where system owners delegate decisions to an automated system. There is scope to forecast, and plan for, some of these unintended harms, given that this is a gap in current cyber-risk management approaches.

For instance, a social media platform may enact an automated decision to remove inflammatory posts, which may result in abrasive users moving to another platform which only becomes an echo chamber that amplifies community division. A service owner might also enact a policy to automatically remove users who fit into a particular profile (such as being based in a particular geographical area), without realizing that this would also catch some legitimate users. This last example also illustrates a moral hazard in cyber-risk management decisions – a practitioner or an automated decision-making system may impose a harm on a subset of users, without experiencing or seeing any comparable harm themselves (such as when a user is excluded from a platform, and in turn has no way to communicate with the platform to make their predicament known). Automated decision-making may impose such harms at a quicker rate than a human service owner; there have already been cases where employees have lost their job to automated decision-making due to none of the humans in the process challenging the decision, even where it was simply because a manager overlooked to renew system access privileges (so it was in turn presumed that the individual must have had their employment terminated). In this sense, automated decisions may not only remove some user decisions, but also impose additional decisions that the user has to make only because a cyber-risk has been introduced into their life from outside.

TU Delft students Designed for Human Autonomy

The Institute encouraged TU Delft students to write their thesis on this important topic. It is crucial that the future engineers and designers critically engage technological development and make sure it fits our human purposes. These projects reflect the ingenuity and ethical commitment of TU Delft students who are rethinking how technology can support—not diminish—our ability to live autonomous, meaningful lives.

Many students explored these questions across diverse domains, the most outstanding ones are highlighted below. For more, please visit our website.

Top 3* of theses on Design for Human Autonomy in the academic year 2024-2025:

Bram Fok Corneel Terwogt, MSc Architecture, Urbanism and Building Sciences (Urbanism), Faculty of Architecture and the Built Environment.

'The Glass Ceiling. Radical Imaginaries of Spatial Justice in Westland's Horticultural Landscape'.

The study critiques socio-environmental inequalities in Westland's high-tech agriculture, exploring spatial justice and autonomy through radical imaginaries and design.

Karlijn Rieky Ine Borremans, MSc Integrated Product Design Faculty of Industrial Design Engineering.

'VAMO: A product-service system for independent visually impaired mobility'.

This project enhances rollators with sensory feedback to support visually impaired users, improving safety, autonomy, and independent mobility. With digital feedback and user-centred design, it aims to create a more adaptive and empowering mobility aid.

Ameya Sawant, MSc Strategic Product Design, Faculty of Industrial Design Engineering.

'On the Nature of Designer Autonomy: Strategic Design Framing Practice in the Age of AI'.

In this thesis, Sawant redefines designer autonomy in the age of AI, exploring how strategic designers can maintain ethical agency in human-machine co-creative processes.

* in random order

→

Honourable mention of theses surrounding Design for Human Autonomy:

Ianthe van Alkemade, MSc Strategic Product Design, Faculty of Industrial Design Engineering.

'Fostering collective self-reliance in elderly care: A value-based approach to the intake conversation in homecare at Surplus'.

This project responds to the growing challenges in Dutch elderly care due to an ageing population. It focuses on uncovering the core human values in receiving and providing care, and explores how to reshape the caregiving role towards shared responsibility.

Aleksander Buszydlik, MSc Computer Science, Faculty of Electrical Engineering, Mathematics and Computer Science

'Finding Recourse for Algorithmic Recourse: Actionable Recommendations in Real-World Contexts'.

This research explores algorithmic recourse in social welfare. It shows how it can support autonomy, encourage engagement, and enable oversight. The study calls for context-specific, multidisciplinary approaches.

5. Conclusions

As was stated in the introduction of this White Paper, human autonomy is not a given: we are not necessarily born with the ability to make choices independently and give reasons why we have made these choices. Technical systems that are based on the assumption that people are autonomous, may affect the human capacity for autonomy, as such systems will disincentivize people to make choices themselves. In the end, human autonomy can only become a reality if the proper conditions are met. Designing for human autonomy means that technical systems empower people to make choices that they can defend as their own.

What the explorations in this white paper show is that designing for human autonomy may be a challenge because autonomy is a heterogeneous concept; however, it is certainly possible and desirable in order to improve the quality of life. What we have shown is that autonomy is a context-dependent concept which needs to be scrutinized and discussed in every design; the White Paper has provided a number of starting points for this.

The introduction presented three nested conceptions of human autonomy, namely agency, freedom, and responsibility. What the different contributions showed is that of these conceptions, agency receives the most attention. Freedom and responsibility seem to be implied in different ways in the design for autonomy. We can see that agency is either secured or facilitated by setting the right design criteria. These two approaches appear to have contrastive ramifications for agency-as-responsibility and agency-as-freedom.

First, securing human autonomy deals with the question of how systems can be designed so that they will not jeopardize the human ability to make independent decisions. Securing autonomy in the context of autonomous systems might imply that we shift from autonomy-as-agency to autonomy-as-responsibility, in the sense that we might reduce the range of choices while we maintain our ability to give good reasons for accepting a specific technical system.

Second, designs can also facilitate human autonomy in the sense that they improve the capacity of individuals to make choices independent from external circumstances. This facilitation of agency can most typically be found in the medical domain, where technical systems help people to make decisions they otherwise could not take. Though such systems cater for individual freedom, they might not be accessible to every person – meaning that some will be left even with less relative autonomy.

Designing for digital autonomy has proven to make our lives more convenient, but as it has been emphasized in this White paper: this should not go to the extent of human autonomy – which is a value for which we can also design.

Critical reflections from the PhD Course: Design for Human Autonomy (10-14 February 2025, TU Delft)

The PhD course *Design for Human Autonomy* brought together researchers and PhD students to critically examine how technology shapes, supports, or undermines human autonomy. Through expert talks—from AI and manipulation to relational autonomy and neuroenhancement the participants got some wonderful insights. The participants of the course wrote after the course a series of insightful blogs, wherein they explored diverse contexts—from AI productivity tools to military health monitoring—revealing the complexity of designing for autonomy.

Several blogs questioned who benefits from automation, for instance in *Automating Autonomy: for whom and by whom?*, while others, like *Siri, how am I?*, critiqued the individualisation of self-monitoring technologies through the lens of relational autonomy. The tension between frictionless design and meaningful engagement was explored in *Giving up on autonomy in the name of autonomy*, highlighting how convenience can paradoxically erode agency.

Other contributions focused on value conflicts and the need for context-sensitive design, such as in *Conflicts when Designing for Autonomy and Is Autonomy Overrated?*, which advocated for vulnerability-based approaches in elder care. The blog posts Lean on me and *Maintaining the Spirit of Human Autonomy* emphasized the importance of supportive relationships and trust in sustaining autonomy.

Innovative perspectives emerged in *AI-Based Clinical Decision Support Systems as a Patient-Power-Up*, which reimagined patient empowerment beyond traditional involvement, and *Tracking Troops*, which examined autonomy under institutional constraints. Together, these reflections underscore that designing for autonomy is not a one-size-fits-all endeavour. It requires ethical sensitivity, contextual awareness, and a willingness to embrace complexity—principles at the heart of this course.

This course was organized in collaboration with the OZSW and 4TU Centre for Ethics and Technology. Read the blogs on the 4TU website: https://www.4tu.nl/ethics/blog/

Literature list

ABB Robotics (2012). “System 800xA Operator Effectiveness Control rooms of the future are here now.” At: https://library.e.abb.com/public/23e1ec050575dc4fc12579e6002b8d22/3BSE068791_en_800xA_Control_rooms_of_the_future_are_here_now.pdf

ABB Robotics (n.d.). “How to enhance control room operator capacities: human factors and ergonomics”, accessed August 25, 2022. At: https://new.abb.com/control-rooms/features/how-to-enhance-control-room-operator-capacities

Agarwal, B., Urlings, C., van Lankveld, G., & Klemke, R. (2022). Ethical FRAPPE–an adapted draft framework for ethical AIED. In Proceedings of: Conference on Technology Enhanced Learning.

Agashe, H. A., Paek, A. Y., & Contreras-Vidal, J. L. (2016). Multisession, noninvasive closed-loop neuroprosthetic control of grasping by upper limb amputees. In Progress in brain research (Vol. 228, pp. 107–128). Elsevier

Anderson, J. (2014). Regimes of autonomy. Ethical Theory and Moral Practice, 17, 355-368.

Autor D. H. (2015). “Why are there still so many jobs? The history and future of workplace automation.” Journal of economic perspectives, 29(3), 3–30

Aylsworth, T., & Castro, C. (2024). Respect for Humanity. In Kantian Ethics and the Attention Economy: Duty and Distraction (pp. 11-65). Cham: Springer International Publishing.

Badrulhisham, F., Pogatzki-Zahn, E., Segelcke, D., Spisak, T., & Vollert, J. (2024). Machine learning and artificial intelligence in neuroscience: A primer for researchers. Brain, Behavior, and Immunity, 115, 470–479. https://doi.org/10.1016/j.bbi.2023.11.005

Bakker, A. B., & Oerlemans, W. G. (2019). Daily job crafting and momentary work engagement: A self-determination and self-regulation perspective. Journal of Vocational Behavior, 112, 417-430.

Beautement, A., Sasse, M. A., & Wonham, M. (2016). Productive Security: A Scalable Methodology for Analysing Employee Security Behaviours. USENIX. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/beautement

Bratton B. H. (2015). The Stack: On Software and Sovereignty. Cambridge, MA: The MIT Press.

Burawoy M. (1979). Manufacturing Consent: Changes in the Labor Process under Monopoly Capitalism. Chicago: The University of Chicago Press

Chatterjee, S., Das, S., Ganguly, K., et al. (2024). Advancements in robotic surgery: Innovations, challenges and future prospects. Journal of Robotic Surgery, 18, 28. https://doi.org/10.1007/s11701-023-01801-w

Chua, J. H., Parkin, S., & Furnell, S. (2020). Identifying Unintended Harms of Cybersecurity Countermeasures. IEEE. https://ieeexplore.ieee.org/abstract/document/9037589

Chua, Y.T. Parkin, S., Edwards, M., Oliveira, D., Schiffner, S., Tyson, G., & Hutchings., A (2019). Identifying Unintended Harms of Cybersecurity Countermeasures. IEEE. https://ieeexplore.ieee.org/abstract/document/9037589

Coggins, T. N. (2022). More work for Roomba? Domestic robots, housework and the production of privacy. Prometheus, 38(1), 98–112. https://www.jstor.org/stable/48676469

Comes, T. (2024). AI for crisis decisions. Ethics and Information Technology, 26(1), 12. https://doi.org/10.1007/s10676-024-09750-0

Crawford K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. New Haven: Yale University Press.

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. Springer. https://doi.org/10.1007/978-1-4899-2271-7.

Farina, D., Vujaklija, I., Brånemark, R., et al. (2023). Toward higher-performance bionic limbs for wider clinical use. Nature Biomedical Engineering, 7, 473–485. https://doi.org/10.1038/s41551-021-00732-x

Fjeld J., N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. SSRN Scholarly Paper. Rochester, NY, 2020. DOI: 10.2139/ssrn.3518482.

Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. OUP Oxford.

Formosa, P. (2021). Robot autonomy vs. human autonomy: Social robots, artificial intelligence (AI), and the nature of autonomy. Minds & Machines, 31, 595–616. https://doi.org/10.1007/s11023-021-09579-2

Friedman, B., & Hendry, D. (2019). Value sensitive design: shaping technology with moral imagination. The MIT Press. Spiekermann, S. (2015). Ethical IT Innovation: A Value-Based System Design Approach. CRC Press.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. Early engagement and new technologies: Opening up the laboratory, 55-95.

Friedrich, O., Racine, E., Steinert, S., Pömsl, J., & Jox, R. J. (2018). An Analysis of the Impact of Brain-Computer Interfaces on Autonomy. Neuroethics, 87(3), 268. https://doi.org/10.1007/s12152-018-9364-9

Frischmann, B., & Selinger, E. (2019). Re-engineering humanity. Cambridge University Press.

Gilbert, F. (2015). A threat to autonomy? The intrusion of predictive brain implants. AJOB Neuroscience, 6(4), 4–11.

Han B. (2017). Psychopolitics: Neoliberalism and New Technologies of Power. London: Verso.

Helms Mills, J., Thurlow, A., & Mills, A. J. (2010). Making sense of sensemaking: the critical sensemaking approach. Qualitative Research in Organizations and Management: An International Journal, 5(2), 182-195. https://doi.org/10.1108/17465641011068857

Henrich, J. (2017). The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter: Princeton University Press.

Herley, C. (2014). More Is Not the Answer. IEEE. https://ieeexplore.ieee.org/abstract/document/6671584

High-Level Expert Group on AI. (2019). Ethics guidelines for trustworthy AI. Brussels: EC.

Hill, T. E. (2000). Respect, pluralism, and justice: Kantian perspectives. Clarendon Press.

Huber, M., Van Boven, L., McGraw, a. P., & Johnson-Graham, L. (2011). Whom to help? Immediacy bias in judgments and decisions about humanitarian aid. Organizational Behavior and Human Decision Processes, 115(2), 283-293. https://doi.org/10.1016/j.obhdp.2011.03.003

Huynh, V., Burger, G., Dang, Q. V., Pelgé, R., Boéris, G., Grizzle, J. W., Ames, A. D., & Masselin, M. (2021). Versatile dynamic motion generation framework: Demonstration with a crutch-less exoskeleton on real-life obstacles at the Cybathlon 2020 with a complete paraplegic person. Frontiers in Robotics and AI, 8, 723780. https://doi.org/10.3389/frobt.2021.723780

ILO (International Labour Organization) (2022). Policy Brief. Sustainable Algorithmic Systems. Geneva. At: https://www.ilo.org/actrav/pubs/WCMS_862207/lang--en/index.htm

Jackson P., Gharavi H. & Klobas J. (2006). "Technologies of the self: Virtual work and the inner panopticon." Information Technology & People, Vol. 19 No. 3: 219-243.

Joque J. (2022). Revolutionary Mathematics: Artificial Intelligence, Statistics, and the Logic of Capitalism. London: Verso.

Kamphorst, B. A. (2012). The primacy of human autonomy: understanding agent rights through the human rights framework. In Proceedings of the 1st workshop on rights and duties of autonomous agents (RDA2) (Vol. 885, pp. 19-24).

Kant I (1788) In: Gregor MJ (ed) Critique of practical reason. Cambridge University Press, Cambridge.

Kauppinen, A. (2011). The social dimension of autonomy. Axel Honneth: Critical Essays, 255-302.

Kavehei, O., Hamilton, T. J., Truong, N. D., & Nikpour, A. (2019). Opportunities for Electroceuticals in Epilepsy. Trends in Pharmacological Sciences, 40(10), 735–746. https://doi.org/10.1016/j.tips.2019.08.001

Kemp, C. C., Edsinger, A., Clever, H. M., & Matulevich, B. (2022). The design of Stretch: A compact, lightweight mobile manipulator for indoor human environments. 2022 International Conference on Robotics and Automation (ICRA), 3150-3157. https://doi.org/10.1109/ICRA46639.2022.9811922

Kirkwood, G., Wilkenfeld, J. N., & Dunbar, N. E. (2022). Exoskeletons and the future of work: Envisioning power and control in a workforce without limits. Human-Machine Communication, 4, 187–206. https://search.informit.org/doi/10.3316/informit.461495764559415

Klaming, L., & Haselager, P. (2013). Did My Brain Implant Make Me Do It? Questions Raised by DBS Regarding Psychological Continuity, Responsibility for Action and Mental Competence. Neuroethics, 6(3), 527–539. https://doi.org/10.1007/s12152-010-9093-1

Klenk, M., & Hancock, J. (2019). Autonomy and online manipulation

Laitinen, A., & Sahlgren, O. (2021). AI systems and respect for human autonomy. Frontiers in artificial intelligence, 4, 705164.

Li, J., Huang, J., Liu, J., & Zheng, T. (2022). Human-AI cooperation: Modes and their effects on attitudes. Telematics and Informatics, 73, 101862.

Lin W., Adey P. & Harris T. (2022). "Dispositions towards automation: Capital, technology, and labour relations in aeromobilities." Dialogues in Human Geography, V (0). At: https://doi.org/10.1177/20438206221112165

Lynch C. R. (2021). "Artificial Emotional Intelligence and the Intimate Politics of Robotic Sociality." Space and Polity, 25(2): 184-201.

Mackenzie, C., and Stoljar, N. (2000). Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self. New York: Oxford University Press.

Maslen, H., & Rainey, S. (2020). Control and Ownership of Neuroprosthetic Speech. Philosophy & Technology. https://doi.org/10.1007/s13347-019-00389-0

Mollick E. R. & Rothbard, N. (2014). "Mandatory fun: Consent, gamification and the impact of games at work." The Wharton School research paper series

Muñoz Sanz V. (2018). "Researching Automated Landscapes." WORK BODY LEISURE, M. Otero Verzier (ed.). Berlin: Hatje Cantz, 103-126.

Neurotechnologies: The Next Technology Frontier | IEEE Brain. (n.d.). Retrieved 12 April 2024, from https://brain.ieee.org/topics/neurotechnologies-the-next-technology-frontier/

OECD (Organisation for Economic Co-operation and Development) (2016). Automation and independent work in a digital economy, Policy Brief on the Future of Work. Paris.

OECD. (2020). Recommendation of the Council on Responsible Innovation in Neurotechnology (OECD Legal Instruments OECD/LEGAL/0457). OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0457

Parkin, S., van Moorsel, A., & Inglesant, P. (2019). Security When it is Welcome: Exploring Device Purchase as an Opportune Moment for Security Behavior Change. NDSS Symposium. https://www.ndss-symposium.org/wp-content/uploads/2019/02/usec2019_02-2_Parkin_paper.pdf

Parkin, S., Redmiles, E., Coventry, L., & Sasse, M.A. (2019). Security When it is Welcome: Exploring Device Purchase as an Opportune Moment for Security Behavior Change. NDSS Symposium. https://www.ndss-symposium.org/wp-content/uploads/2019/02/usec2019_02-2_Parkin_paper.pdf

Paulus, D., Fathi, R., Fiedrich, F., de Walle, B. V., & Comes, T. (2022). On the Interplay of Data and Cognitive Bias in Crisis Information Management. Information Systems Frontiers. https://doi.org/10.1007/s10796-022-10241-0

Perrig L. (2021). “Manufacturing consent in the gig economy.” Augmented Exploitation. Artificial Intelligence, Automation and Work, P. Moore, J. Woodcock (eds.). London: Pluto Press, 75–86.

Pesch, U. (2024). Values, Institutions and Innovations for Societal Progress: The Moral Pursuit of Autonomy and Rationality. Cheltenham and Northampton: Edward Elgar.

Pesch, U. (2024). Values, Institutions and Innovations for Societal Progress: The Moral Pursuit of Autonomy and Rationality. Edward Elgar Publishing

Prendergast, J. M., Balvert, S., Driessen, T., Seth, A., & Peternel, L. (2021). Biomechanics aware collaborative robot system for delivery of safe physical therapy in shoulder rehabilitation. IEEE Robotics and Automation Letters, 6(4), 7177-7184. https://doi.org/10.1109/LRA.2021.3097375

Prunkl C., “Human autonomy in the age of artificial intelligence”. In: Nature Machine Intelligence 4.2 (2022), pp. 99–101. DOI: 10/gsd2rt.

Rainey, S. (2023). A gap between reasons for skilled use of BCI speech devices and reasons for utterances, with implications for speech ownership. Frontiers in Human Neuroscience, 17, 1248806. https://doi.org/10.3389/fnhum.2023.1248806

Raz, J. (2001). Value, Respect, and Attachment. Cambridge: Cambridge University Press.

Read J. (2024). The Double Shift. Spinoza and Marx on the Politics of Work. London: Verso.

Reeder, R. W., Ion, I., Consolvo, S., & Sleeper, M. (2017). 152 Simple Steps to Stay Safe Online: Security Advice for Non-Tech-Savvy Users. IEEE. https://ieeexplore.ieee.org/abstract/document/8055663

Rejeb, A., Rejeb, K., Simske, S., & Treiblmaier, H. (2021). Humanitarian Drones: A Review and Research Agenda. Internet of Things, 16, 100434. https://doi.org/https://doi.org/10.1016/j.iot.2021.100434

Riley, S., and Bos, G. (2021). Human Dignity. Internet Encyclopedia of Philosophy. Available at: www.iep.utm.edu/hum-dign/ (accessed May 2, 2021).

Ripstein, A. (1999). Equality, Responsibility, and the Law. Cambridge: Cambridge University Press.

Rubel, A., Castro, C., and Pham, A. (2020). Algorithms, Agency, and Respect for Persons. Social Theory and Practice, 46 (3), 547–572. doi:10.5840/soctheorpract202062497

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American psychologist, 55(1), 68.

Ryan, R. M., & Deci, E. L. (2004). Autonomy is no illusion. Handbook of experimental existential psychology, 455.

Ryan, R. M., & Deci, E. L. (2006). Self-regulation and the problem of human autonomy: Does psychology need choice, self-determination, and will?. Journal of personality, 74(6), 1557-1586.

Ryan, R. M., & Deci, E. L. (2017). Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. New York: Guilford Press.

Sadowski J. (2022). “Planetary Potemkin AI: The Humans Hidden inside Mechanical Minds”, Digital Work in the Planetary Market, M. Graham, F. Ferrari (Eds.). Cambridge, MA: The MIT Press.

Sandvik, K., Jacobson, K., & McDonald, S. (2017). Do no harm: A taxonomy of the challenges of humanitarian experimentation. ICRC Review, forthcoming
Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction, 36(6), 495-504.

Song, H., Hsieh, T. H., Yeon, S. H., et al. (2024). Continuous neural control of a bionic limb restores biomimetic gait after amputation. Nature Medicine, 30, 2010–2019. https://doi.org/10.1038/s41591-024-02994-9

Suchman L., “The uncontroversial ‘thingness’ of AI”. In: Big Data & Society 10.2 (2023), pp. 1–5. DOI: 10/gs6q9w.

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. Internet policy review, 8(2), 1-22.

Taskin L., Parmentier, M., & Stinglhamber, F. (2019). “The dark side of office designs: towards de-humanization.” New Technology, Work and Employment, 34(3), 262-284.

Taylor A. (2018). “The Automation Charade,” Logic Magazine, no. 5 (August). At: https://logicmag.io/failure/the-automation-charade

Tech Tesseract, “TRANSFORM your GAMING SETUP with RGB Lighting! How I light my Gaming Room!”, YouTube video, accessed August 25, 2022, https://www.youtube.com/watch?v=dZgABJ72y20

Tribelhorn, B., & Dodds, Z. (2007). Evaluating the Roomba: A low-cost, ubiquitous platform for robotics research and education. Proceedings 2007 IEEE International Conference on Robotics and Automation, 1393-1399. https://doi.org/10.1109/ROBOT.2007.363179

Vallor, S. (2015). Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. Philosophy & Technology, 28(1), 107-124. https://doi.org/10.1007/s13347-014-0156-9

Van Grunsven, J., & Roeser, S. (2022). AAC technology, autism, and the empathic turn. Social Epistemology, 36(1), 95-110.

Van de Poel, I. (2013). Translating values into design requirements. In D. Mitchfelder, N. McCarty, & D. E. Goldberg (Eds.), Philosophy and Engineering: Reflections on Practice, Principles and Process (pp. 253-266). Springer.; page 261

Wakefield, J. (2018, June 21). The man who was fired by a machine. BBC News. https://www.bbc.com/news/technology-44561838

Walker, M. J., & Mackenzie, C. (2020). Neurotechnologies, Relational Autonomy, and Authenticity. IJFAB: International Journal of Feminist Approaches to Bioethics, 13(1), 98–119. https://doi.org/10.3138/ijfab.13.1.06

Williams, N. R., & Okun, M. S. (2013). Deep brain stimulation (DBS) at the interface of neurology and psychiatry. The Journal of Clinical Investigation, 123(11), 4546–4556.

Wu, M. H., Yu, J. C., & Lin, Y. C. (2022). Study of autonomous robotic lawn mower using multi-sensor fusion based simultaneous localization and mapping. 2022 International Conference on Advanced Robotics and Intelligent Systems (ARIS), 1-4. https://doi.org/10.1109/ARIS56205.2022.9910445

Zaga, C., Lupetti, M. L., Forster, D., Murray-Rust, D., Prendergast, M., & Abbink, D. (2024). First international workshop on worker-robot relationships: Exploring transdisciplinarity for the future of work with robots. HRI 2024. https://doi.org/10.1145/3610978.3638156

de Mul, J., & van den Berg, B. (2011). Remote control: Human autonomy in the age of computer-mediated agency. In Law, human agency and autonomic computing (pp. 46-63). Routledge.

van Nes, I. J. W., van Dijsseldonk, R. B., van Herpen, F. H. M., Rijken, H., Geurts, A. C. H., & Keijsers, N. L. W. (2022). Improvement of quality of life after 2-month exoskeleton training in patients with chronic spinal cord injury. The Journal of Spinal Cord Medicine, 47(3), 354–360. https://doi.org/10.1080/10790268.2022.2052502

van Vliet, M. H. M., Versluis, A., Chavannes, N. H., Scheltinga, B. L., Albers, N., Penfornis, K. M., Baccinelli, W., Meijer, E., & on behalf of the Perfect Fit consortium (2024). Protocol of a mixed-methods evaluation of Perfect Fit: A personalized mHealth intervention with a virtual coach to promote smoking cessation and physical activity in adults. Digital Health, 10. https://doi.org/10.1177/20552076241300020

Van Vliet, M., De Kleijn, M., van den Brekel-Dijkstra, K., Huijts, T., van Hogen-Koster, S., Jung, H. P., & Huber, M. (2024, March). Rapid review on the concept of positive health and its implementation in practice. In Healthcare (Vol. 12, No. 6, p. 671). MDPI.

van Wynsberghe, A., & Comes, T. (2020). Drones in humanitarian contexts, robot ethics, and the human–robot interaction. Ethics and Information Technology, 22(1), 43-53.

van de Poel, I. (2015). Conflicting Values in Design for Values. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), Handbook of Ethics, Values, and Technological Design (pp. 89-116). Springer Netherlands. https://doi.org/10.1007/978-94-007-6994-6_5-1

van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. Minds and machines, 30(3), 385-409.

Notes

Colofon

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

Projectleader: Ivo Kostons

Graphic design: Jet van Zwieten

Fonts: Lars Medium, by Bold Decisions / Minion Pro Medium, by Robert Slimbach.

Printing: BK Prints

Website: www.delftdesignforvalues.nl
Email: info@delftdesignforvalues.nl

Part of Annual Theme Design for Human Autonomy
White Paper Delft Design for Values Institute, 2025