Meaningful human control: actionable properties for AI system development

Cavalcante Siebert, L.; Lupetti, M.L.; Aizenberg, E.; Beckers, N.W.M.; Zgonnikov, A.; Veluwenkamp, H.M.; Abbink, D.A.; Giaccardi, Elisa; Houben, G.J.P.M.; Jonker, C.M.

DOI
10.1007/s43681-022-00167-3

Publication date
2022

Document Version
Final published version

Published in
AI and Ethics

AI and Ethics
https://doi.org/10.1007/s43681-022-00167-3

ORIGINAL RESEARCH

Meaningful human control: actionable properties for AI system development

Luciano Cavalcante Siebert[1,2] · Maria Luce Lupetti[1,3] · Evgeni Aizenberg[1,2] · Niek Beckers[1,4] · Arkady Zgonnikov[1,4] · Herman Veluwenkamp[1,5] · David Abbink[1,4] · Elisa Giaccardi[1,3] · Geert-Jan Houben[1,2] · Catholijn M. Jonker[1,2] · Jeroen van den Hoven[1,5] · Deborah Forster[1] · Reginald L. Lagendijk[1,2]

Received: 7 January 2022 / Accepted: 15 April 2022

Abstract
How can humans remain in control of artificial intelligence (AI)-based systems designed to perform tasks autonomously? Such systems are increasingly ubiquitous, creating benefits - but also undesirable situations where moral responsibility for their actions cannot be properly attributed to any particular person or group. The concept of meaningful human control has been proposed to address responsibility gaps and mitigate them by establishing conditions that enable a proper attribution of responsibility for humans; however, clear requirements for researchers, designers, and engineers are yet inexistent, making the development of AI-based systems that remain under meaningful human control challenging. In this paper, we address the gap between philosophical theory and engineering practice by identifying, through an iterative process of abductive thinking, four actionable properties for AI-based systems under meaningful human control, which we discuss making use of two applications scenarios: automated vehicles and AI-based hiring. First, a system in which humans and AI algorithms interact should have an explicitly defined domain of morally loaded situations within which the system ought to operate. Second, humans and AI agents within the system should have appropriate and mutually compatible representations. Third, responsibility attributed to a human should be commensurate with that human's ability and authority to control the system. Fourth, there should be explicit links between the actions of the AI agents and actions of humans who are aware of their moral responsibility. We argue that these four properties will support practically minded professionals to take concrete steps toward designing and engineering for AI systems that facilitate meaningful human control.

Keywords Artificial intelligence · AI ethics · Meaningful human control · Moral responsibility · Socio-technical systems

1 Introduction

Deploying AI algorithms in human-inhabited environments comes with the risk of inappropriate, undesirable, or unpredictable consequences [1, 2]. The misinterpreted skills of AI systems, combined with their rapid impact in public and private spheres of life, can lead to situations with a clear misalignment between human moral values and societal norms [3–6], and where moral responsibility for such undesired impacts can often not be properly attributed to any person [7].

How can designers, users, or other human agents be morally responsible for systems that are designed to perform tasks, learn, and adapt without direct human control? The strong technical drive towards achieving systems which can act independently from human control in more scenarios, does not necessarily include considerations about

Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov: Co-first authors.

✉ Luciano Cavalcante Siebert
L.CavalcanteSiebert@tudelft.nl

1 AiTech Interdisciplinary Research Program on Meaningful Human Control, Delft University of Technology, Delft, The Netherlands

2 Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

3 Faculty of Industrial Design Engineering, Delft University of Technology, Delft, The Netherlands

4 Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Delft, The Netherlands

5 Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands

Published online: 18 May 2022

the socio-technical consequences of implementing these systems, especially in terms of facilitating moral responsibility. In agreement with the hybrid intelligence community [8], we believe a stronger focus on human-AI systems (systems in which humans and AI algorithms interact during operation) is needed to address the complex design issue of ethical use and implementation of AI [9]. The very features that make AI algorithms useful complicate their assessment and predictability in the complex socio-technical context in which they operate - which changes over time. As a result, all systems based on AI, especially those with so-called higher "levels of autonomy" ,[1] can and should be designed for appropriate human responsibility [11]. The holy grail is to design these systems in a manner that can mitigate the occurrence of situations that the manufacturer was in principle unable to anticipate, and that users were not able to appropriately influence or even realize.

The problem of designing for human responsibility over human-AI systems is challenging because such systems operate in complex social infrastructures that include organizational processes with both human-to-human and human-AI interactions, policy, and law. Designing for moral responsibility therefore requires a systemic, socio-technical perspective that jointly considers the interaction between all these elements [12]. This fundamental challenge of intertwined social, physical, and technical infrastructures does not exclusively concern AI: societies have settled on morally acceptable solutions for ubiquitous technology in other domains, such as medicine and aviation safety.

However, these solutions do not readily generalize to systems based on AI algorithms, due to properties such as: (1) learning abilities; (2) black-box nature; (3) impact on many stakeholders (even those not using the systems themselves); and (4) autonomous or semi-autonomous decision-making features.

First, AI agents can demonstrate novel behavior through learning from historical data and continuous learning via interactions with the world and other agents. Because the world we are concerned with is an open system with respect to the the agents' perceptions and actions, the behavior of human-AI systems cannot be predicted with precision over time [13, 14]. Second, the agent's decision-making process may be difficult to explain and predict, even for its programmer [15], complicating responsibility attribution for its consequences. Third, as AI agents may interact with multiple users, which have different levels of expertise, different preferences, and understanding, responsibility can become a diffuse concept for which no one feels morally engaged. This may be further exacerbated when AI agent's autonomous features are overestimated by those interacting with it. As the system's design process may overlap with implementation and use [16], interactions may end up including humans who did not choose to be involved in its use, as in the case of sidewalk pedestrians interacting with automated vehicles. Fourth, as systems based on AI with increasing autonomous decision-making features operate with reduced or even no meaningful supervision, undesirable impacts might be perceived only in hindsight. Learning abilities, opacity, interaction with many stakeholders, and autonomous or semi-autonomous features are just four of the prominent issues, which emerge as algorithms interact with social environments.

To design for moral responsibility and human control is particularly important as quick development and immediate deployment "in the wild" [17], instead of regulated tests procedures, is urging academia and governments to take a stance in defining visions for trustworthy AI [5]. In fact, even if the "move fast and break things" mantra was considered acceptable and received wide consensus for driving digital innovation in the last decade, the same cannot be for AI with autonomous features [18]. A failure of an AI agent is not a "404 error page". It is a car accident, most likely with fatalities [19, 20]; it is an unfair and discriminatory distribution of wealth and services [21]; it is an unjust crime accusation based on ethnicity [22, 23]. Designers and developers of AI systems can only tackle this challenge by acknowledging upfront that successful attribution and apportioning of responsibility is not a matter of fortuitous allocation of praise or blame.

The concept of *meaningful human control* [11, 24–26] was first proposed to address the problem of responsibility gaps in autonomous weapon systems, but is becoming a central concept when discussing responsible AI [11].The core idea is that humans should ultimately remain in control of, and thus morally responsible for, the behavior of human-AI systems.[2] Nevertheless, meaningful human control has also received the critique to be an ill-defined concept [29] that ignores operational context [30] and does not provide concrete design guidelines [12].

This article aims to contribute to closing the gap between the theory of meaningful human control, as proposed by

[1] "Level of autonomy" is a complex construct. In line with Bradshaw's seven deadly myths of autonomy [10], we acknowledge that measuring autonomy on a single ordered scale of increasing levels is insufficient because it lacks context, is not human-centred, and disregard functional differences, among other reasons.

[2] Meaningful human control relates not only to the engineering of the AI agent, but also to the design of the socio-technical environment that surrounds it, including social and institutional practices [11, 27, 28]. As [12] elaborate, "[intelligent] devices themselves play an important role but cannot be considered without accounting for the numerous human agents, their physical environment, and the social, political and legal infrastructures in which they are embedded."

[11], and the practice of designing and developing human-AI systems by proposing four actionable properties that can be addressed throughout the system's lifecycle. We start by unpacking the philosophical concept of meaningful human control (Sect. 2). We then present a set of four properties that were generated through an iterative process of abductive thinking that combined the different disciplinary perspectives of the authors (engineering, computer science, philosophy of technology, ethnography and design). We describe each property and illustrate how each of them helps defining whether and to what extent a human-AI system is under meaningful human control. We also suggest concrete methods and tools that can support addressing each property and illustrate them with respect to two case studies: automated vehicles and AI-based hiring (Sect. 3). Finally, we discuss the systemic and socio-technical nature of these properties and the need for transdisciplinary practices (Sect. 4) and conclude the paper (Sect. 5).

2 Meaningful human control: tracking and tracing

The concept of meaningful human control was coined in the debates on autonomous weapon systems [24, 25]. At the heart of this concept is the idea that humans need to retain control and moral responsibility over autonomous systems. This discussion is no longer exclusive to the military domain. Meaningful human control is increasingly relevant in other domains as AI agents become more ubiquitous and autonomous, especially in non-forgiving scenarios in which fundamental human rights and safety are at stake. The concept has already been discussed on the context of automated vehicles [12, 31, 32], including truck platooning [33], surgical robots [34], smart home systems [35], medical diagnosis [36], and content moderation in social media [37].

Although many authors agree on the need for some form of human control over AI agents [11, 24, 25, 30], these same authors may diverge and often disagree about what makes human control meaningful. Observing the theoretical challenges of specifying what meaningful human control means, Santoni de Sio and Van den Hoven [11] laid out a foundation for a theory of meaningful human control with an adaptation of Fischer and Ravizza's [38] philosophical account on guidance control, moral responsibility, and free will. Following the ideals of responsible innovation [39] and value-sensitive design [40], a centerpiece of Santoni de Sio's and Van den Hoven's conception of meaningful human control is two necessary conditions for meaningful human control, tracking and tracing:

(1) *Tracking* condition: to be under meaningful human control, a human-AI system should be responsive to the human moral reasons relevant in the circumstances. A human-AI system that fulfills this condition is said to track the relevant human moral reasons.
(2) *Tracing* condition: in order for a human-AI system to be under meaningful human control, its behavior, capabilities, and possible effects in the world should be traceable to a proper moral and technical understanding on the part of at least one relevant human agent who designs or interacts with the system.

In the tracking condition, control is said to be meaningful when the system's performance co-varies with the reasons of the relevant person or persons, like a mercury column in a thermometer co-varies with the temperature in the room. When air humidity varies, but the temperature remains constant, we expect no change in the mercury column, since it only tracks temperature. Similarly, when someone always accepts a new job only because the salary is higher, that person tracks financial gain, not necessarily the job's intrinsic reward.

For the tracing condition, it is required that this relevant person or persons are in a position to have a proper moral and technical understanding of the system. That would not be the case if the thermometer would randomly induce changes in the mercury column, or if the concept of the mercury expanding or contracting is entirely unknown to the person. Similarly, a supervisor of an automated vehicle that does not understand traffic rules, would not have such understanding.

The tracking and tracing conditions take the concept of meaningful human control one step closer to support practical design and development because they provide high-level design requirements for a human-AI system to be under meaningful human control. Building on this conception, researchers developed frameworks to analyze and quantify factors affecting meaningful human control for automated vehicles [12, 41, 42]. However, a description of general system-level properties that could support operationalization of tracking and tracing conditions in diverse contexts is yet to be specified.

3 Four properties of human-AI systems under meaningful human control

The tracking and tracing conditions [11] provide a philosophical grounding for informing the development of human-AI systems under meaningful human control. Yet, translating these philosophical concepts into a concrete design and engineering practice is far from trivial. For instance, the tracking condition suggests that a human-AI system should be responsive to the moral reasons of a relevant human. But, *how do we define the relevant human in*

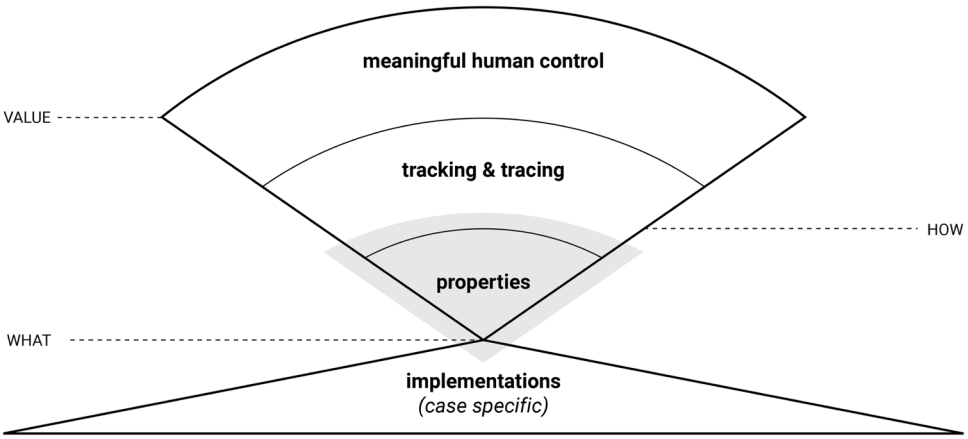

Fig. 1 The diagram, based on the framework of abductive thinking by Dorst [45], illustrating the positioning of the abductive thinking we performed in search for a solution space that would meet the claimed need for meaningful human control

a given circumstance? How should a given AI system recognize a moral reasoning? Does the condition imply that every AI system should be designed to be morally sensitive [43]*?* The tracing condition implies the necessity of a proper moral and technical understanding from at least one relevant human interacting and designing the system. *Does this imply that the AI system should be able to recognize if and when an interacting human has such proper moral and technical understanding?* Or, *does this imply that we need protocols for the design and use of AI systems that define if and when a human can and must have such understanding?*

In an effort to answer these questions — and more — the authors, a group of researchers from various backgrounds (engineering, computer science, philosophy of technology, ethnography and design), engaged in an iterative process of abductive thinking [44]. Specifically, we built on Dorst's conceptual framework of abductive thinking [45], where both a desired value (meaningful human control) and a working principle (tracking & tracing conditions) are known, to brainstorm ideas of what the solutions to achieve these might be. The generated ideas were then grouped into thematic areas and synthesized into actionable properties. It has to be noted, however, that although this work explores the solution space of the framework, our aim is rather to provide a contribution that sits on a meta-level, in between the what and the how (see Fig. 1).

Specifically, as a result of our abductive thinking where we collectively reflected on what strategic and engineering solutions would enable the two necessary conditions of meaningful human control [11], we identified a set of four actionable properties. In the following subsections we describe these properties in detail and illustrate their practical implications. To better highlight the properties and the implications, we make use of two example application scenarios: automated vehicles[3] and AI-based hiring[4]. Both cases manifest an urgent need for meaningful human control in non-forgiving scenarios that strongly impact people's

[3] *Automated vehicle scenario:* A conditionally automated vehicle that can perform operational aspects of the driving task (e.g., lane keeping or adaptive cruise control) as well as tactical aspects: detecting events and objects on the road and responding to them, and interacting with human pedestrians and other vehicles. Under normal circumstances, the automated vehicle can complete a whole trip without interventions from the human driver; the manufacturer emphasizes these autonomous features in their marketing and promotional materials. The driver, however, is required to constantly supervise the system. There is no requirement for the driver to keep their hands on the steering wheel, but the driver must remain alert at all times and be able to take over operational control at the request of the automation system. The vehicle does not actively monitor the driver state, but in case a human intervention is required, it attracts the driver's attention through a visual alert message and a loud auditory signal. The automated driving system used in the vehicle relies on machine learning-based object recognition and behavior prediction components, which were trained on the data obtained during extensive testing on public roads.

[4] *AI-based hiring scenario:* Job candidates applying for a vacancy go through an automated video interview where they record their answers to questions formulated ahead of time by the employer. After the interview is completed, an AI agent applies machine learning methods to quantify candidates' suitability for the job by correlating their facial expressions, choice of words, and voice tone to personal traits such as creativity, willingness to learn, and conscientiousness. To tailor the AI agent towards the context-specific preferences of the employer, the machine learning algorithms were trained on video interviews performed with current employees and their respective annual performance evaluations. The employer sets a threshold for a passing score, and based on the scores outputted by the AI agent, a list of candidates who pass to the next selection round is automatically compiled. The candidates do not see the score they were assigned. Neither the candidate, nor the employer, receive an explanation of how the scores were computed. The employer considers the human-AI system to be a cost-effective solution for what has previously been a time-consuming first-round selection process that required hiring additional screening staff. In addition, the employer seeks to increase diversity at the company and considers AI-based selection to be less prone to discriminatory biases.

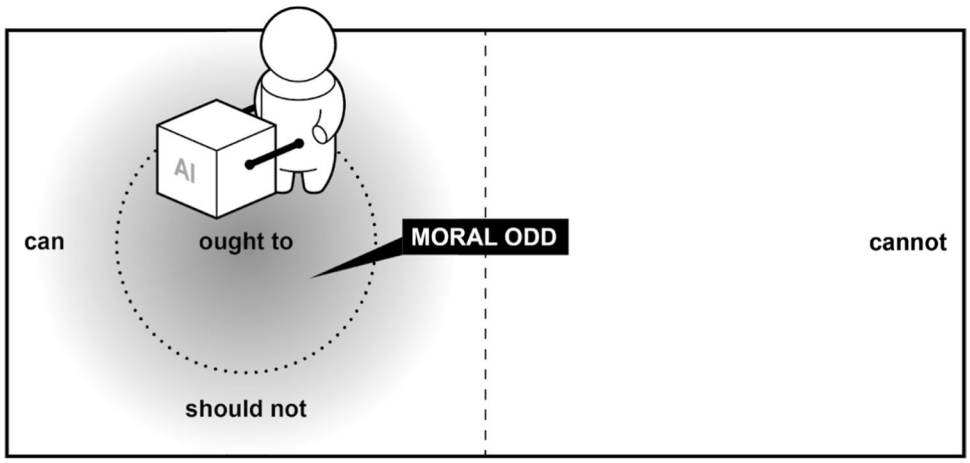

Fig. 2 Property 1: Moral ODD. The human-AI system should operate within the boundaries of what it can do (for both the human and the AI agent) and within the moral boundaries of what it ought to do, i.e. the human-AI system should act according to the relevant moral reasons of the relevant stakeholders

lives (e.g., bodily harm, unfair decisions, discrimination), and their differences with respect to time constraints, embodiment, and involved stakeholders juxtapose different aspects of realizing these properties in human-AI systems.

3.1 Property 1. The human-AI system has an explicit moral operational design domain (moral ODD) and the AI agent adheres to the boundaries of this domain

As the human-AI system has to be "responsive to relevant human moral reasons" (i.e., the tracking condition), we need to identify the relevant humans, their relevant (moral) reasons, and the circumstances in which these reasons are relevant. To this end, specifying the technical conditions in which the system is designed to operate is not sufficient. Designers should consider a larger design space, one that captures also the values and societal norms that must be considered and respected during both design and operation.

Building on the concept of operational design domain (ODD) which originates in the automotive domain [46], we name this larger design space the *moral operational design domain* (moral ODD). The concept of ODD is often used in the context of automated driving and refers to a set of contextual conditions under which a driving automation system is designed to function: outside of it a human driver is responsible. Specific contextual properties of the automotive ODD typically include factors like road structure, road users, road obstacles and environmental conditions (material elements), as well as human-vehicle interactions and expected vehicle interactions with pedestrians (relational elements) [47]. In terms of legal responsibility, the ODD constitutes a selection of operation scenarios that can be safely managed [48] by the automation and in which undesired consequences are minimized [47]. As such, we believe it is a valuable concept to extend beyond automated vehicles, but for human-AI systems in general.

The current conceptualization of the ODD strongly focuses on the technical aspects of operation and the goal to extend the context boundaries of the ODD. However, consideration of the wider societal implications is lacking. Similar to Burton et al. [49], we argue that the concept of ODD should also emphasize the broader social and ethical implications. We propose this extended concept of ODD so that functional considerations of where and when a human-AI system *can* operate, are seconded and complemented to the definition of the domain in which a system is *ought* or *should not* operate from a moral perspective (Fig. 2).

A simple example of a hammer illustrates the difference between the "*can*" (e.g., material and relational elements) and the "*ought to*" dimensions (e.g., moral elements). From a purely functional perspective, a hammer "can" be used as a weapon against another person. However, the morally acceptable use of a hammer is for hammering nails (ought to), not to injure other people (should not). Common sense already tells us that the use of a hammer as a weapon is in most cases morally unacceptable (can but should not). It is clear that the responsibility for proper use lies with the user, not the manufacturer (except in cases where the hammer clearly does not function properly, e.g., the head suddenly comes loose from the handle and injures a person).

In a scenario involving complex human-AI systems, this is often much less clear cut. In the automated vehicle case, the moral ODD could contain moral reasons representing safety (e.g., avoid road accidents), efficiency (e.g., reduce travel time), and personal freedom (e.g., enhance independence for seniors), to name just a few. In the AI-based hiring context, moral reasons could include, from the employer's side, reducing discrimination or increasing the number of applicants in the recruitment process, while for the applicants' side autonomy over self-representation could be considered very relevant. In both contexts, however, there might

be tensions among different moral reasons and stakeholders, requiring an inclusive specification and careful communication of the moral ODD.

3.1.1 Practical considerations

The specification and clear communication of the moral ODD support relevant humans (e.g., users, designers, developers) to be aware of the moral implications of the system's actions and their responsibility for these actions, thereby supporting the tracing condition of meaningful human control. Furthermore, if the operation of the AI agent remains confined within the boundaries of what it "can do" and "ought to do", the tracking condition of meaningful human control is supported as well, as this makes the human-AI system more responsive to human understanding of what is the morally appropriate domain and mode of operation. Achieving these benefits requires that: (1) the moral ODD be explicitly defined; (2) the AI agent embed concrete solutions to constrain the actions of the human-AI system within the boundaries of the ODD.

To define the moral ODD, designers and developers need to engage with fundamental questions of what are the elements composing the moral ODD and how do the features of each element affect the system's behavior. The process starts with an ontological modelling of the environment(s) in which the human-AI system is expected to operate. Such complex assemblage of elements and relationships could be meaningfully represented within the moral ODD by making use of principles from existing research on software applications where ontologies are developed to enable context-aware computing systems [50, 51]. The mapping of material and relational elements characterizing a domain should be complemented with an investigation of what might be the morally relevant reasons, what they represent in the specific context, assumptions and consequences related to the system operation. Such understanding of the moral landscape of an AI agent under development could be built by means of extensive literature and case reviews [52–54], participatory approaches such as interviews, interactive workshops, and value-oriented coding of qualitative responses [40], which can be supported by natural language processing algorithms [55].

How to satisfy the second requirement (constraining the AI agent to the boundaries of the moral ODD) varies according to the constituent elements of the moral ODD. When constraining the material and relational aspects of the system behavior, approaches developed in the automotive and aircraft domains can be a useful reference, e.g., risk-based path planning strategies for unmanned aircraft systems in populated areas [56] and geofencing [57]. Relational aspects can be addressed through envelope protection. In the aircraft domain, flight envelope protection systems prevent the pilot from making control commands that drive the aircraft outside its operational boundaries, a concept that has also been adopted for unmanned aerial vehicles [58]. This concept could be extended beyond the aircraft domain, and become a more general design pattern for constraining the relational elements of the moral ODD in the systems involving both embodied and non-embodied AI agents [59].

Moral constraints are arguably the most challenging to enforce. One possible way of imposing them is to set probabilistic guarantees on system outcomes [60]. However, these approaches might not hold in real-world applications. Due to the non-quantifiable nature of morally relevant elements, as well as moral disagreements among humans, the boundaries of the moral ODD will remain blurred [49]. Hence, it is crucial that humans, not AI agents, are empowered to be aware of their responsibilities to make conscious decisions if and when the human-AI system should deviate from the boundaries defined by the moral ODD. The assessment of whether and how an AI agent is confined to the moral ODD is not a binary check, but rather a contextualized and deliberated analysis of the interaction between the AI agent, human agents, and the social, physical, ethical, and legal environment surrounding them. Humans, to conclude, should have an understanding of such blurry boundaries of the moral ODD and their responsibility to meaningfully control the AI agent in this process. Importantly, this includes the possibility of deciding that the use of an AI agent is not acceptable in certain contexts.

3.2 Property 2. Human and AI agents have appropriate and mutually compatible representations of the human-AI system and its context

For a human-AI system to perform its function, both humans and AI agents within the system should have some form of representations of the involved tasks, role distributions, desired outcomes, the environment, mutual capabilities and limitations. Such representations are often referred to as *mental models*; these models enable agents to describe, explain and predict the behavior of the system and decide which actions to take [61–63].

Shared representations, i.e., representations that are mutually compatible between human and AI agents within the system, allow the agents to have appropriate understanding of each other, the task, and the environment [62], which facilitates agents to cooperate, adapt to changes, and respond to relevant human reasons. To ensure safe operation of the system, agents should also have a shared representation of each other's abilities and limitations. Specifically, the AI agents should account for humans' inherent physical and cognitive limitations, while human agents should account for the AI agents' limitations to avoid issues such

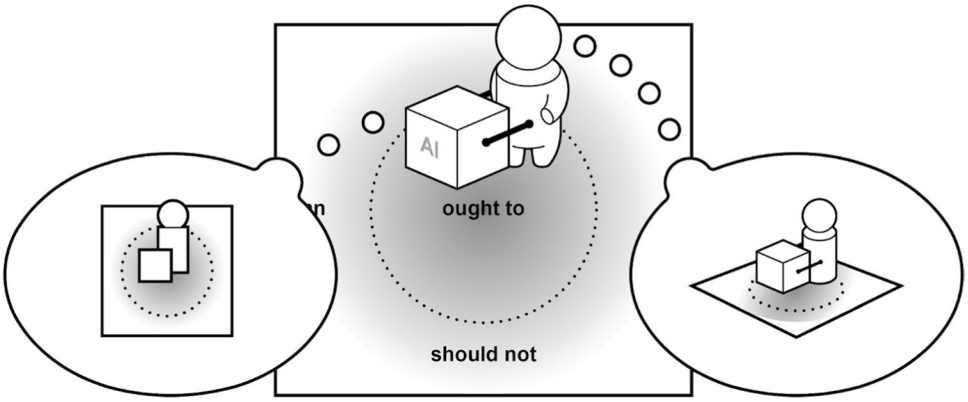

Fig. 3 Property 2: The human and AI agents have appropriate and mutually compatible representations of the human-AI system and of each other's abilities and boundaries

as overreliance [64]. Furthermore—crucial to achieve meaningful human control—these shared representations should include the human reasons identified in the moral ODD (Fig. 3), which can change over time and across contexts. Due to the dynamic nature of elements of the shared representations, the human and AI agents should be able to update their representations of the potentially changing reasons accordingly.

Incompatibility between representations could result in the lack of responsiveness to human reasons, thereby leading to undesired outcomes with significant moral consequences. For example, inconsistent mental models between a human driver and automated vehicle about "who has the control authority", in which the human driver believes that the automated vehicle has control and vice versa, could result in a critical and unsafe system state [65].

3.2.1 Practical considerations

In order for the agents' shared representations to facilitate the system's tracking of relevant human reasons, the system designers first need to define which aspects of the system and its context (including relevant humans, AI agents, the environment, and the moral ODD) each agent should have a representation of. The process of determining what kinds of representations are needed will be context-specific and depend on the moral ODD of the system. A useful approach to determine the necessary representations and to translate these high-level concepts into practical design requirements is co-active design [61]. Specific to building and maintaining shared representations, this approach provides guidelines on how to establish observability and predictability between the human and AI agents, including what needs to be communicated and when [62].

Representations can include practical matters such as task allocation, role distribution and system limits, but also understanding of how humans perceive the AI agents, human acceptance of and trust in the human-AI system, humans values and social norms. This should also include determining the appropriate level of representation. For instance, for an automated vehicle to interact with a pedestrian, the designers need to determine whether it suffices for the vehicle to have a representation of just the location of a pedestrian on the road and their movement trajectory, or also the height and age of that human, their goals and intentions. In the context of AI-based hiring, a key aspect requiring shared representation is the meaning of competence. In particular, the meanings of soft skills, such as teamwork and creativity, are highly fluid, context-dependent, and contestable. Therefore, aligning the job-specific meaning of competence among job seekers, employers, and any AI agent involved in the hiring process is critical.

Once the representations required for each agent are defined, the design and engineering choices need to sufficiently take these into account. Specifically, such choices should facilitate (1) AI agents to build and maintain representations of the humans and their reasons, and (2) humans to form mental models of AI agents and the overall human-AI system. These shared representations can be achieved through various combinations of implicit (e.g., through interaction between agents) or explicit ways (e.g., by means of human training, verbal communication). For example, to allow humans to build and maintain a representation of an AI agent, it can be developed to be observable and predictable implicitly through its design (e.g., glass-box design [66]), allowing the operator to better understand the AI agent's decision-making. Ecological interface design can also leverage knowledge on human information processing to design human-AI interfaces that are optimally suited to convey complex data in a comprehensible manner [67]. Maintaining accurate representations during the human-AI system's deployment can also occur through interaction, either implicitly (e.g., through intent inference from observed behavior) or explicitly (e.g., explicit verbal or written messages). For example, an AI system can probe through behavior whether the human is aware of its intentions before committing to a decision [68].

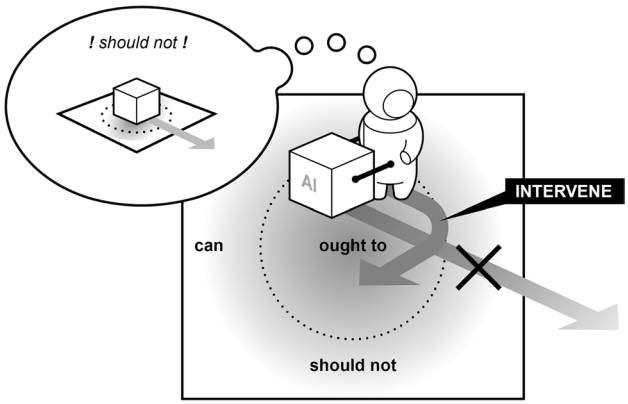

Fig. 4 Property 3: The relevant humans and AI agents have the ability and authority to control the system so that humans can act upon their responsibility, e.g., if the human recognizes that a given situation might bring the system outside the moral ODD, they can intervene to avoid this

For the AI agents to have appropriate representations of human agents, the assumptions about human intentions and behavior adopted by AI agents (either implicitly or explicitly) need to be validated. This can be aided by incorporating theoretically grounded and empirically validated models of humans in the interaction-planning algorithms of AI agents [69, 70], or by augmenting bottom-up, machine-learned representations with top-down symbolic representations [71, 72]. An alternative approach, value alignment [73], aims to mitigate the problems that arise when autonomous systems operate with inappropriate objectives. In particular, inverse reinforcement learning (IRL), which is often used in value alignment, aims to infer the "reward function" of an agent from observations of the agent's behavior, also in cooperative partial-information settings (cooperative IRL) [74]. Although IRL is likely not sufficient to infer human preferences from observed behaviour since human planning systematically deviates from the assumed global rationality [75], such approaches could still support agents to maintain aligned shared representations [76].

3.3 Property 3. The relevant humans and AI agents have ability and authority to control the system so that humans can act upon their responsibility

Relevant humans should not be considered just mere subjects to be blamed in case something goes wrong, i.e., an ethical or legal scapegoat for situations when the system goes outside the moral ODD. They should rather be in a position to act upon their moral responsibility by influencing the AI system throughout its operation, and to bring the system back to the moral ODD if needed (Fig. 4).

This is only possible when the distribution of roles and control authority between humans and AI ("who is doing what and who is in charge of what") is consistent with their individual and combined abilities, including reasonable mechanism for overruling the AI agent through intervening and correcting behavior, setting new goals, or delegating sub-tasks.

Flemisch et al. [65] provide a thorough account on the importance of an appropriate balance between an agent's ability, authority, and responsibility in human–machine systems: ability to control should not be smaller than control authority, and control authority should not be smaller than responsibility. We argue that this account applies to complex human-AI systems as well. The *ability* of a human or AI agent includes their skill and competence to perceive the state of a system and the environment. This also includes a way to acquire and analyze relevant information, to make a decision to act, and to perform that action appropriately [77]. Ability also includes the resources at their disposal, such as tools (an autonomous vehicle without a steering wheel would severely hamper the human's ability to control the vehicle's direction; job candidates' ability to represent themselves would be heavily impaired by the lack of a feedback mechanism) or time (an automated vehicle that would wait until the very last second to alert the driver of a dangerous situation also limits the driver's ability to direct the vehicle to safety; an employer would have no control and understanding of an AI-based hiring system if assessment of candidates would be provided only after the selection process finishes).

The understanding of an (AI or human) agent's *ability* is intrinsically related to the socio-technical context in which the system is embedded. Hence, it is important that tasks are distributed according to the agent's ability in the context, not only from a functional perspective, but also accounting for the values and norms intrinsic to the activity. Approaches such as the nature-of-activities [78, 79], under the umbrella of Value Sensitive Design [40], can support the understanding of which set of tasks should be (partially or totally) delegated or shared with AI agents, and which should be left exclusively to humans. Given the collaborative nature of many human-AI systems, team design patterns can be used as an intuitive graphical language for describing and communicating to the team the design choices that influence how humans and AI agents collaborate [80, 81].

The second component of the account proposed in [65] is control *authority*, i.e., the degree to which a human or AI agent is enabled to execute control. Consistency between authority and ability requires that an agent's authority does not exceed their ability. And similarly, responsibility should not exceed authority. Thus, an agent should be responsible only for tasks they have authority to perform, and they

should have authority only over tasks they are able to perform. A key implication of this consistency is that control is exerted by the agent that has sufficient ability and authority, and more responsibility is carried by the agents that exert more control. While ability and authority are attributes that both human and AI agents possess, we consider responsibility as a human-only quality. Therefore, the ability and authority of a human-AI system must be traced to responsibilities of relevant humans, e.g., engineers, designers, operators, users, and managers.

In the automated vehicle case, the driver has authority to control the vehicle by accelerating, breaking and steering, as well to take over control authority at any time. In the case of AI-based hiring, employers' authority includes setting a threshold for a passing score and deciding who to hire. Simply giving human agents final authority by design, without ensuring proper ability, is not sufficient to empower humans to act upon their moral responsibility. For example, a driver may have final authority over a fully autonomous car, but the driver's loss of situational awareness, or even skill degradation as a result of systematic lack of engagement in the driving task, will limit the driver's ability to exert that control authority [41, 65, 82]. The same might happen for a manager with final authority over who to hire, if they merely sign off on the hiring recommendations of the AI agent, without substantively engaging in the assessment process themselves.

3.3.1 Practical considerations

As *authority* should not be smaller than *ability*, it is important to build a baseline understanding of the abilities of human and AI agents and evaluate their consistency with the control authority provided by the system's design. From the human side, human factors literature [83] can support the identification of a realistic baseline on human ability by applying psychological and physiological principles to understand challenges that are likely to arise in human-AI interaction [82, 84]. From the AI side, a proper understanding of ability should not only be task-oriented (e.g., measuring performance from data sets against benchmark), but also behavior-oriented. Approaches to understand AI ability in context include approaches inspired by human cognitive tests, information theory [85], and ethology (related to animal behavior) [14]. Designing for appropriate authority and ability also requires us to expand the scope of design from human-AI interactions to social and organizational practices [11]. Human training, oversight procedures, administrative discretion, and policy are just a few examples of organizational elements that significantly determine and shape agents' authority and ability.

Design, training and technological development may "expand" or "shrink" agents' abilities through innovation, including training humans for new skills and equipping AI agents with new technological capabilities, or achieving more through interaction between humans and AI and their combined abilities. From the AI side, especially for machine learning-based systems, as the relation between the input data and the target variable changes over time, concept drift methodologies can be applied to identify new situations which might impact the AI agent's ability to respond to new situations [86, 87]. From the human side, interaction with technology might lead to behavioral adaptation and unwanted situations, e.g., speeding when driving with intelligent steering assistance provided by an automated vehicle [88], decreasing human's ability to keep the system within the moral ODD. In such situations, the human-AI system might move to a fallback state [89] or attract the driver's attention back to the supervision task thus restoring the driver's ability to act upon their ultimate responsibility for the vehicle's operation.

Shared control is a promising approach to keep a balance between control ability and authority, with relevant applications in the domain of automated vehicles, robot-assisted surgery, brain-machine interfaces, and learning [90]. In shared control, the human(s) and the AI agents(s) are interacting congruently in a perception-action cycle to perform a dynamic task, i.e., control authority is not attributed either to the human or to the AI agent, but is shared among them [91]. Shared control could be particularly useful in human-AI systems that need to act in complex situations that can rapidly change beyond the envisioned moral ODD, and where rapid human adaptation and intervention is needed.

3.4 Property 4. Actions of the AI agents are explicitly linked to actions of humans who are aware of their moral responsibility

Satisfying the first three properties ensures that relevant humans are capable of acting upon their moral responsibility (property 3), are aware of the moral implications of the system's actions (property 1), and have shared representations with AI agents (property 2). Yet, what is left undiscussed is the requirement to ensure that the effects of the system's actions are traceable to the relevant humans' moral understanding.

To trace any consequence of the human-AI system's operation to a proper moral understanding of relevant humans, there should be explicit, explainable and inspectable link(s) between actions of the system and corresponding human morally loaded decisions and actions. We acknowledge that such link(s) might be a more demanding form of tracing than what was originally proposed in [11], nevertheless we deem it necessary to enable the tracing condition to be inspectable. Furthermore, we argue that moral understanding of the system's effects should be demonstrated by, at least, those

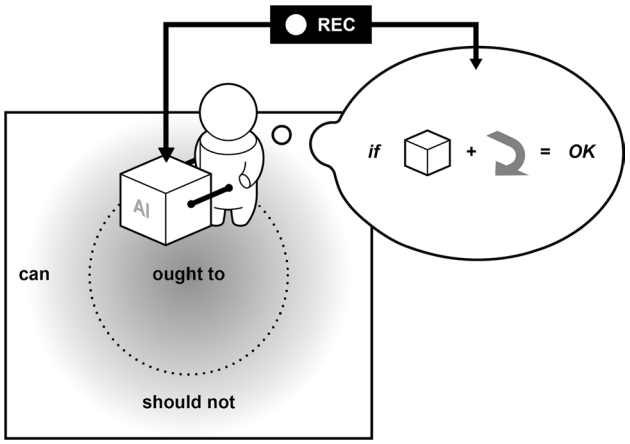

Fig. 5 Property 4: Actions of the AI agents are explicitly linked to actions of humans who are aware of their moral responsibility. The forward link starts on the human and indicates that whenever a human makes a decision with moral implications that affects the system's behavior, that human should be aware of their moral responsibility. The backward link looks at actions of the human-AI system and links it to previous human decisions (e.g., designers, users)

humans who make decisions with moral implications on the design, deployment, or use of the system, even if the actions that bring a human decision to life are executed by the AI agent. Hence, all relevant human decisions related to e.g., design, use, policy must be explicitly logged and reported [66], in order to link actions of the AI agents to relevant decisions, preferences, or actions of humans who are aware of the system's possible effects in the world.

Even if all relevant humans made their decisions responsibly and with full awareness of their possible moral implications, the lack of a readily identifiable link from a given action of the AI agent to the underlying human decisions would still result in loss of tracing. The links between actions of the systems and corresponding human morally loaded decisions and actions need to be explicitly identifiable in two ways (Fig. 5):

(1) *Forward link*: whenever a human within the human-AI system makes a decision with moral implications (e.g., on the design, deployment, or use of the system), that human should be aware of their moral responsibility associated with that decision, even if the actions that bring this decision to life are executed by the AI agent.
(2) *Backward link*: for any consequence of the actions of the human-AI system, the human decisions and actions leading to that outcome should be readily identifiable.

3.4.1 Practical considerations

Enabling the *forward link* from human moral understanding to AI agents' actions relates to the epistemic condition (also called knowledge condition) of moral responsibility, which posits that humans should be aware of their responsibility at the time of a decision [3, 92]. Hence, the human-AI system should be designed in a way that simplifies and aids achieving moral awareness. This requires explicit links between design choices and stakeholder interpretations of moral reasons that are at stake. Values hierarchies [93] provide a structured and transparent approach to map relations between design choices and normative requirements. A value hierarchy visualizes the gradual specification of broad moral notions, such as moral responsibility, into context-dependent properties or capabilities the system should exhibit, and further into concrete socio-technical design requirements. Such a structured mapping can equip stakeholders with the means to deliberate design choices in a manner that explicitly links each choice to relevant aspects of moral responsibility. These deliberations, as well as the accompanying rich body of empirical and conceptual research must be well documented, inspectable, and legible. This kind of transparency also supports the *backward link* between the system's actions and the design choices made by relevant humans.

Furthermore, requirements such as explainability of the system's actions can be essential in effectively empowering human moral awareness. Since its early works, the field of explainable AI has increased its scope from explaining complex models to technical experts towards placing the target audience as a key aspect [94]. Given a certain human or group of humans as target audience, we see explainability in the context of supporting the forward link as clearly presenting the link between the system's actions and human moral awareness, as well as their alignment to the moral ODD. For example, consider an automated vehicle which slows down and pulls off the road after it recognizes a car accident [12]. Right after that the vehicle should then remind the driver of their duty to provide assistance to possible victims in the accident. In the context of AI-based hiring, explanations of assessment scores in language that directly links observed job seeker performances to job-specific meanings of competence can help employers, job seekers, designers, and developers better outline the boundaries of moral ODD during design phase. For example, this can help reveal whether there is misalignment between conceptions of competence among the human agents and the AI algorithm.

In complex socio-technical environments the establishment of links between human moral awareness and actions of a human-AI system is complicated by the "problem of many hands", which happens when more than one agent contributes to a decision. It becomes less clear who is morally and legally responsible for its consequences [95]. The "problem of many things" complicates this further: there are not only many (human) hands, but also many different technologies interacting and influencing each others, be it multiple AI agents or the interplay between sensors,

processing units, and actuators [96]. In case of unintended consequences of the AI agent's actions, this complexity can hinder the backward link, i.e., tracing the responsibility back to individual human decisions. This challenge calls for systemic, socio-technical design interventions that jointly consider social infrastructure (e.g., organizational processes, policy), physical infrastructure, and the AI agents that are part of these infrastructures.

Recent developments using information theory to quantify human causal responsibility [97] can provide relevant insight for the design and development of appropriate forward and backward links, by providing a model with which hypotheses can be tested. However, simplifying assumptions used in this research need to be addressed to account for more realistic settings. Methods from social sciences, e.g., Actor-Network Theory (ANT) [98] can support the development of tracing networks of association amongst many actors, which can help understand how, for example, humans may offload value-laden behavior onto the technology around us. In the "sociology of a door closer", [99] describes how we made door closers the element in the assembly that manifests politeness by ensuring the door closes softly and gradually, even as the human actors may barge through without any action to regulate the door). This sort of division of moral-labor should not be done mindlessly, it requires human decisions to be analyzed and their relation to the moral ODD to be carefully analyzed.

Although establishing explicit links between human decisions, human moral awareness, and actions of the AI agents is challenging, they allow appropriate post hoc attribution of backward-looking responsibility for unintended consequences, helping to avoid responsibility gaps and prevent similar events from repeating in the future. It also facilitates forward-looking responsibility by creating an incentive for the relevant humans to proactively reflect on the consequences of their decisions (design choices, operational control, interactions, etc.).

3.5 Summary of the four properties

We summarize the proposed properties of systems under meaningful human control as follows:

- *Property 1:* The human-AI system has an explicit moral operational design domain (moral ODD) and the AI agent adheres to the boundaries of this domain.
- *Property 2:* Human and AI agents have appropriate and mutually compatible representations of the human-AI system and its context.
- *Property 3:* The relevant agents have ability and authority to control the system so that humans can act upon their responsibility.
- *Property 4:* Actions of the AI agents are explicitly linked to actions of humans who are aware of their moral responsibility.

In our view, these properties are constructive as well as open: they can serve as practical tools for supporting the design, development and evaluation of human-AI systems, while being applicable to diverse types of systems (as illustrated by the cases of automated vehicles and AI-based hiring).

Although the properties are not sufficient for a system to be under meaningful human control, we deem them necessary from a design perspective: while a system developed to possess all these properties may still not be fully under meaningful human control, we believe that completely missing one of these properties would imply that the human-AI system is not under meaningful human control. Moreover, each property is non-binary and necessarily multidimensional. Consequently, improving the system to some extent according to one or more of the four properties will lead to better tracking or tracing, and therefore, more meaningful human control over that system. That said, defining "how much of each property is sufficient" in a given context would generally require a thorough qualitative and situated analysis.

Furthermore, these four properties in themselves do not immediately translate to concrete design guidelines; metrics, algorithms, and methodologies needed to implement the properties are context- and system-specific. Yet, the properties provide explicit anchors for connecting to existing frameworks and methodologies across the design and engineering domains.

4 The broader picture

In addition to establishing explicit links between the concept of meaningful human control and existing frameworks across the design and engineering domains, the four proposed properties unveil a range of new methodological questions and challenges on the path to practically implementing systems under meaningful human control.

Designing for meaningful human control requires designing for emergence. We argue that improving the human-AI system according to the properties we presented will lead to better tracking and tracing, and therefore more meaningful human control over that system. However, that does not provide an answer to the critical question: how much meaningful human control is sufficient in a given context? We believe these uncharted waters need to be explored through practice-based research that aims to responsibly develop human-AI systems, while

ensuring inclusive and transparent collaborations among stakeholders and safe and rigorous evaluation of concepts and designs. On one hand, it is reasonable to expect that socio-technical design requirements that act for the sake of meaningful human control properties will vary across societal and application domains. On the other hand, given a sufficient level of conceptual abstraction, a common basic set of system properties that will prove practically helpful and robust across different societal domains can inform both bottom-up practice and top-down regulation towards meaningful human control. However, design and regulation cannot account for every detail of a system's processes, interactions, components in a deterministic, top-down fashion. In fact, the socio-technical complexity of human-AI systems and the inherent uncertainty of some aspects of their operation call for designing for emergence [100], where the focus shifts to designing the social, physical, and technical infrastructures that jointly provide favorable conditions for interactions between agents to lead to emergence of desirable system properties and behaviors.

Meaningful human control is necessary but not sufficient for ethical AI. Meaningful human control over AI relates to the broader scope of AI ethics in the sense that designing for meaningful human control means designing for human moral responsibility. That is a critical aspect of ethical design of human-AI systems, but by itself it is not sufficient to ensure other crucial aspects of ethical design and operation, such as protection of human rights and environmental sustainability. It is possible for a human-AI system to be under meaningful human control with respect to some relevant humans, yet result in outcomes that are considered morally unacceptable by society at large [11]. Meaningful human control ensures that humans are aware of and are equipped to act upon their responsibility, and that the human-AI system is responsive to human moral reasons. But it does not prevent humans from consciously designing and operating the human-AI system in an unethical way. Therefore, meaningful human control must be part of a larger set of design objectives that collectively align the human-AI system with societal values and norms.

Transdisciplinary practices are vital to achieve meaningful human control over AI. One of the most prominent challenges threaded throughout the four properties may also be the most rewarding opportunity: the inherent need for a socio-technical design process that crosses disciplinary boundaries. Each of the four properties and meaningful human control as a whole is an endeavor that is not solvable by a single discipline. It is a systemic, socio-technical puzzle in which computer scientists, designers, engineers, social scientists, legal practitioners, and crucially, the societal stakeholders in question, each hold an essential piece of the puzzle. Hence, the only way to "walk the walk" is to move forward together, forming a transdisciplinary practice based on continuous mutual learning [101] among both academic and non-academic stakeholders. While this is undoubtedly a challenge, it may prove to be a rewarding opportunity for socially inclusive innovation that puts human moral responsibility front and center.

5 Conclusion

In this article, we address the issue of responsibility gaps in design and use of AI systems, and argue in favor of the concept of meaningful human control as a principle to mitigate them. To the current discourse surrounding meaningful human control, we contribute with a set of four actionable system properties and related approaches useful for implementing them in practice. These properties unpack the tracking and tracing conditions of meaningful human control [11] and provide a significant step forward toward its operationalization. Even though these properties may not be sufficient to completely ensure meaningful human control for all possible situations, we deem them necessary, and as such they help translate the tracking and tracing conditions into more tangible and designable requirements for human-AI systems. Our properties build upon and expand existing conceptual frameworks and methodologies across the design and engineering domains, such as the notion of operational design domain [47], ontological modeling [51], co-active design [61], shared mental models [62], shared control [90], value alignment [73], and consistency of ability, control authority, and responsibility [65]. With these four properties we have realized two goals: (1) contributed to closing the gap between the philosophical theory and practice of designing systems under meaningful human control, and (2) explicitly link meaningful human control to existing frameworks and methodologies across disciplines that can support design and development of human-AI systems.

Societal impacts and the issue of responsibility gaps in the use of AI today puts forward meaningful human control as one of the central concepts when discussing trustworthy and responsible AI, and we think it should also take central place on AI development. We believe this work will enable researchers and practitioners to take actionable steps towards the design and development of systems under meaningful human control, enabling many of the promised benefits of AI while maintaining human responsibility and control.

Acknowledgements We thank Filippo Santoni de Sio for helpful comments on the earlier version of this manuscript.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Floridi, L., Cowls, J., King, T.C., Taddeo, M.: How to design AI for social good: seven essential factors. Sci. Eng. Ethics **26**(3), 1771–1796 (2020). https://doi.org/10.1007/s11948-020-00213-5
2. Stinson, C.: Algorithms are not neutral. AI and Ethics, 1–8 (2022)
3. Coeckelbergh, M.: AI Ethics. The MIT press essential knowledge series, The MIT Press, Cambridge, MA (2020)
4. Cruz, J.: Shared Moral Foundations of Embodied Artificial Intelligence. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 139–146 (2019)
5. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
6. Umbrello, S., De Bellis, A.F.: A value-sensitive design approach to intelligent agents. In: Roman Y (ed.), Artificial Intelligence Safety and Security (2018) CRC Press (2018)
7. Matthias, A.: The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf. Technol. **6**(3), 175–183 (2004). https://doi.org/10.1007/s10676-004-3422-1
8. Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., Van der Gaag, L., van Harmelen, F., Van Hoof, H., Van Riemsdijk, B., van Wynsberghe, A., Verbrugge, R., Verheij, B., Vossen, P., Welling, M.: A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer **53**(8), 18–28 (2020). https://doi.org/10.1109/MC.2020.2996587
9. Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., Génova, G., Haim, G., Kließ, M.S., Lopez-Sanchez, M., Micalizio, R., Pavón, J., Slavkovik, M., Smakman, M., Van Steenbergen, M., Tedeschi, S., Van der Toree, L., Villata, S., de Wildt, T.: Ethics by Design: Necessity or Curse? In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 60–66. ACM, New Orleans LA USA (2018). https://doi.org/10.1145/3278721.3278745
10. Bradshaw, J.M., Hoffman, R.R., Woods, D.D., Johnson, M.: The seven deadly myths of "autonomous systems''. IEEE Intell. Syst. **28**(3), 54–61 (2013)
11. Santoni de Sio, F., Van den Hoven, J.: Meaningful human control over autonomous systems: a philosophical account. Front. Robot. AI **5**, 15 (2018)
12. Mecacci, G., Santoni de Sio, F.: Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. Ethics Inf. Technol. **22**(2), 103–115 (2020). https://doi.org/10.1007/s10676-019-09519-w
13. Johnson, N., Zhao, G., Hunsader, E., Qi, H., Johnson, N., Meng, J., Tivnan, B.: Abrupt rise of new machine ecology beyond human response time. Sci. Rep. **3**(1), 2627 (2013). https://doi.org/10.1038/srep02627
14. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O., Jennings, N.R., Kamar, E., Kloumann, I.M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D.C., Pentland, A.S., Roberts, M.E., Shariff, A., Tenenbaum, J.B., Wellman, M.: Machine behaviour. Nature **568**(7753), 477–486 (2019)
15. European Parliamentary Research Service (EPRS): The ethics of artificial intelligence: Issues and initiatives. Panel for the Future of Science and Technology PE 634.452 (2020)
16. Giaccardi, E., Redström, J.: Technology and more-than-human design. Design Issue **36**(4) (2020)
17. Chen, L., Wilson, C.: Observing algorithmic marketplaces in-the-wild. ACM SIGecom Exchang **15**(2), 34–39 (2017). https://doi.org/10.1145/3055589.3055594
18. Taplin, J.: Move Fast and Break Things: How Facebook, Google, and Amazon Have Cornered Culture and What It Means For all of us. Pan Macmillan, New York (2017)
19. Johnston, P., Harris, R.: The Boeing 737 MAX saga: lessons for software organizations. Softw. Qua. Profession. **21**(3), 4–12 (2019)
20. Serter, B., Beul, C., Lang, M., Schmidt, W.: Foreseeable Misuse in Automated Driving Vehicles-The Human Factor in Fatal Accidents of Complex Automation. Technical Report 0148-7191, SAE Technical Paper (2017)
21. Korinek, A., Stiglitz, J.E.: Artificial intelligence and its implications for income distribution and unemployment. Technical Report 0898-2937, National Bureau of Economic Research (2017)
22. Angwin, J.: Jeff Larson, Surya Mattu, Lauren Kirchner: Machine Bias There's software used across the country to predict future criminals. ProPublica, And it's biased against blacks (2016)
23. Sweeney, L.: Discrimination in online ad delivery. Queue **11**(3), 10–29 (2013). https://doi.org/10.1145/2460276.2460278
24. Article 36: Key areas for debate on autonomous weapons systems: Memorandum for delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS) (2014)
25. Article 36: Killing by Machine: Key Issues for Understanding Meaningful Human Control (2015)
26. Horowitz, M.C., Scharre, P.: Meaningful Human Control in Weapon Systems: A Primer. working paper, 15 (2015)
27. Behymer, K.J., Flach, J.M.: From autonomous systems to socio-technical systems: designing effective collaborations. She Ji: J. Des. Econ. Innovat. **2**(2), 105–114 (2016)
28. Santoni de Sio, F., Mecacci, G.: Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philos. Technol. pp. 1–28 (2021)
29. Cummings, M.: Lethal autonomous weapons: meaningful human control or meaningful human certification? IEEE Technol. Soc. Magn. **38**(4), 20–26 (2019). https://doi.org/10.1109/MTS.2019.2948438
30. Ekelhof, M.: Moving beyond semantics on autonomous weapons: meaningful human control in operation. Global Pol. **10**(3), 343–348 (2019). https://doi.org/10.1111/1758-5899.12665
31. Beckers, G., Sijs, J., Van Diggelen, J., van Dijk, R.J.E., Bouma, H., Lomme, M., Hommes, R., Hillerstrom, F., Van der Waa, J., Van Velsen, A., Mannucci, T., Voogd, J., Van Staal, W., Veltman, K., Wessels, P., Huizing, A.: Intelligent autonomous vehicles with an extendable knowledge base and meaningful human control. In: Bouma, H., Stokes, R.J., Yitzhaky, Y., Prabhu, R. (eds.) Counterterrorism, Crime Fighting, Forensics, and Surveillance

Technologies III, p. 11. SPIE, Strasbourg, France (2019). doi: 10.1117/12.2533740
32. Calvert, S.C., Heikoop, D.D., Mecacci, G., Van Arem, B.: A human centric framework for the analysis of automated driving systems based on meaningful human control. Theor. Issues Ergonom. Sci. 1–29 (2019). https://doi.org/10.1080/1463922X.2019.1697390
33. Calvert, S.C., Mecacci, G., Heikoop, D.D., de Sio, F.S.: Full platoon control in Truck Platooning: A Meaningful Human Control perspective. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 3320–3326. IEEE, Maui, HI (2018). https://doi.org/10.1109/ITSC.2018.8570013
34. Ficuciello, F., Tamburrini, G., Arezzo, A., Villani, L., Siciliano, B.: Autonomy in surgical robots and its meaningful human control. Paladyn. J. Behav. Robot. **10**(1), 30–43 (2019)
35. Umbrello, S.: Meaningful human control over smart home systems: a value sensitive design approach. Hum. Mente J. Philos. Stud. **12**(37) (2020)
36. Braun, M., Hummel, P., Beck, S., Dabrock, P.: Primer on an ethics of AI-based decision support systems in the clinic. J. Med. Ethics 2019–105860 (2020). https://doi.org/10.1136/medethics-2019-105860
37. Wagner, B.: Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. Policy Internet **11**(1), 104–122 (2019)
38. Fischer, J.M., Ravizza, M.: Responsibility and control: a theory of moral responsibility. Cambridge University Press, Cambridge (1998)
39. Van den Hoven, J.: Value Sensitive Design and Responsible Innovation. In: Owen, R., Bessant, J., Heintz, M. (eds.) Responsible Innovation, pp. 75–83. Wiley, Chichester, UK (2013). https://doi.org/10.1002/9781118551424.ch4
40. Friedman, B., Hendry, D.G.: Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press, Cambridge (2019)
41. Heikoop, D.D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., Van Arem, B.: Human behaviour with automated driving systems: a quantitative framework for meaningful human control. Theor. Issues Ergon. Sci. **20**(6), 711–730 (2019). https://doi.org/10.1080/1463922X.2019.1574931
42. Calvert, S.C., Mecacci, G.: A conceptual control system description of Cooperative and Automated Driving in mixed urban traffic with Meaningful Human Control for design and evaluation. IEEE Open Journal of Intelligent Transportation Systems, 1–1 (2020). https://doi.org/10.1109/OJITS.2020.3021461
43. Wallach, W., Allen, C., Smit, I.: Machine morality: bottom-up and top-down approaches for modelling human moral faculties. In: Machine Ethics and Robot Ethics, pp. 249–266. Routledge (2020)
44. Timmermans, S., Tavory, I.: Theory construction in qualitative research: from grounded theory to abductive analysis. Sociol. Theory **30**(3), 167–186 (2012)
45. Dorst, K.: The nature of design thinking. In: Design Thinking Research Symposium (2010). DAB Documents
46. SAE: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Technical report, SAE International (2018). https://doi.org/10.4271/J3016_201806
47. Czarnecki, K.: Operational Design Domain for Automated Driving Systems - Taxonomy of Basic Terms (2018)
48. Koopman, P., Fratrik, F.: How Many Operational Design Domains, Objects, and Events? 4 (2019)
49. Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., Porter, Z.: Mind the gaps: assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. Artif. Intell. **279**, 103201 (2020). https://doi.org/10.1016/j.artint.2019.103201
50. Bettini, C., Brdiczka, O., Henricksen, K., Indulska, J., Nicklas, D., Ranganathan, A., Riboni, D.: A survey of context modelling and reasoning techniques. Pervasive Mob. Comput. **6**(2), 161–180 (2010). https://doi.org/10.1016/j.pmcj.2009.06.002
51. Cabrera, O., Franch, X., Marco, J.: 3LConOnt: a three-level ontology for context modelling in context-aware computing. Softw. Syst. Model. **18**(2), 1345–1378 (2019). https://doi.org/10.1007/s10270-017-0611-z
52. Coeckelbergh, M.: Drones, information technology, and distance: mapping the moral epistemology of remote fighting. Ethics Inf. Technol. **15**(2), 87–98 (2013). https://doi.org/10.1007/s10676-013-9313-6
53. Galliott, J.: Military Robots: Mapping the Moral Landscape. Ashgate Publishing Ltd, London (2015)
54. Childress, J.F., Faden, R.R., Gaare, R.D., Gostin, L.O., Kahn, J., Bonnie, R.J., Kass, N.E., Mastroianni, A.C., Moreno, J.D., Nieburg, P.: Public health ethics: mapping the terrain. .J. Law Med. Ethics **30**(2), 170–178 (2002). https://doi.org/10.1111/j.1748-720X.2002.tb00384.x
55. Liscio, E., Van der Meer, M., Siebert, L.C., Jonker, C.M., Mouter, N., Murukannaiah, P.K.: Axies: Identifying and evaluating context-specific values. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, pp. 799–808 (2021)
56. Primatesta, S., Scanavino, M., Guglieri, G., Rizzo, A.: A Risk-based Path Planning Strategy to Compute Optimum Risk Path for Unmanned Aircraft Systems over Populated Areas. In: 2020 International Conference on Unmanned Aircraft Systems (ICUAS), pp. 641–650. IEEE, Athens, Greece (2020). https://doi.org/10.1109/ICUAS48674.2020.9213982
57. Maiouak, M., Taleb, T.: Dynamic Maps for Automated Driving and UAV Geofencing. IEEE Wirel. Commun. **26**(4), 54–59 (2019). https://doi.org/10.1109/MWC.2019.1800544
58. Yavrucuk, I., Prasad, J.V.R., Unnikrishnan, S.: Envelope Protection for Autonomous Unmanned Aerial Vehicles. J. Guid. Control. Dyn. **32**(1), 248–261 (2009). https://doi.org/10.2514/1.35265
59. Robbins, S.: AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines. AI & Soc. **35**(2), 391–400 (2020). https://doi.org/10.1007/s00146-019-00891-1
60. Thomas, P.S., da Silva, B.C., Barto, A.G., Giguere, S., Brun, Y., Brunskill, E.: Preventing undesirable behavior of intelligent machines. Science **366**(6468), 999–1004 (2019)
61. Johnson, M., Bradshaw, J.M., Feltovich, P.J., Jonker, C.M., Van Riemsdijk, M.B., Sierhuis, M.: Coactive design: designing support for interdependence in joint activity. J. Hum.-Robot Interact. **3**(1), 43 (2014). https://doi.org/10.5898/JHRI.3.1.Johnson
62. Jonker, C.M., Van Riemsdijk, M.B., Vermeulen, B.: Shared Mental Models. In: De Vos, M., Fornara, N., Pitt, J.V., Vouros, G. (eds.) Coordination, Organizations, Institutions, and Norms in Agent Systems VI, pp. 132–151. Springer, Berlin (2011)
63. Wilson, J.R., Rutherford, A.: Mental models: theory and application in human factors. Hum. Factors **31**(6), 617–634 (1989)
64. Lee, J.D., See, K.A.: Trust in Automation: Designing for Appropriate Reliance. Human Factors, 31 (2004)
65. Flemisch, F., Heesen, M., Hesse, T., Kelsch, J., Schieben, A., Beller, J.: Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations. Cognit. Technol. Work **14**(1), 3–18 (2012)
66. Aler Tubella, A., Theodorou, A., Dignum, F., Dignum, V.: Governance by glass-box: implementing transparent moral bounds for AI behaviour. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, pp. 5787–5793. International Joint Conferences on Artificial Intelligence

Organization, Macao, China (2019). https://doi.org/10.24963/ijcai.2019/802
67. Vicente, K.J., Rasmussen, J.: Ecological interface design: theoretical foundations. IEEE Trans. Syst. Man Cybern. **22**(4), 589–606 (1992). https://doi.org/10.1109/21.156574
68. Sadigh, D., Landolfi, N., Sastry, S.S., Seshia, S.A., Dragan, A.D.: Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state. Auton. Robot. **42**(7), 1405–1426 (2018). https://doi.org/10.1007/s10514-018-9746-1
69. Schürmann, T., Beckerle, P.: Personalizing human-agent interaction through cognitive models. Front. Psychol. **11**, 8 (2020). https://doi.org/10.3389/fpsyg.2020.561510
70. Siebinga, O., Zgonnikov, A., Abbink, D.: Validating human driver models for interaction-aware automated vehicle controllers: A human factors approach. arXiv:2109.13077 [cs] (2021)
71. Van Bekkum, M., de Boer, M., Van Harmelen, F., Meyer-Vitali, A., ten Teije, A.: Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases. arXiv:2102.11965 [cs] (2021)
72. Marcus, G.: The next decade in AI: four steps towards robust artificial intelligence. arXiv:2002.06177 [cs] (2020)
73. Gabriel, I.: Artificial intelligence, values, and alignment. Mind. Mach. **30**(3), 411–437 (2020)
74. Hadfield-Menell, D., Russell, S.J., Abbeel, P., Dragan, A.: Cooperative inverse reinforcement learning. Adv. Neural Inf. Process. Syst. **29** (2016)
75. Armstrong, S., Mindermann, S.: Occam's razor is insufficient to infer the preferences of irrational agents. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc., (2018). https://proceedings.neurips.cc/paper/2018/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf
76. Peysakhovich, A.: Reinforcement Learning and Inverse Reinforcement Learning with System 1 and System 2. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. AIES '19, pp. 409–415. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3306618.3314259
77. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man. Cybern. Part A: Syst. Hum. **30**(3), 286–297 (2000)
78. Santoni de Sio, F., Robichaud, P., Vincent, N.A.: Who should enhance? Conceptual and normative dimensions of cognitive enhancement. HUMANA.MENTE J. Philos. Stud. **7**(26), 179–197 (2014)
79. Santoni de Sio, F., Van Wynsberghe, A.: When should we use care robots? The nature-of-activities approach. Sci. Eng. Ethics **22**(6), 1745–1760 (2016)
80. Van Diggelen, J., Johnson, M.: Team Design Patterns. In: Proceedings of the 7th International Conference on Human-Agent Interaction, pp. 118–126. ACM, Kyoto Japan (2019). https://doi.org/10.1145/3349537.3351892
81. Van der Waa, J., Van Diggelen, J., Cavalcante Siebert, L., Neerincx, M., Jonker, C.: Allocation of Moral Decision-Making in Human-Agent Teams: A Pattern Approach. In: Harris, D., Li, W.-C. (eds.) Engineering Psychology and Cognitive Ergonomics. Cognition and Design vol. 12187, pp. 203–220. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-49183-3_16
82. Kyriakidis, M., de Winter, J.C.F., Stanton, N., Bellet, T., Van Arem, B., Brookhuis, K., Martens, M.H., Bengler, K., Andersson, J., Merat, N., Reed, N., Flament, M., Hagenzieker, M., Happee, R.: A human factors perspective on automated driving. Theor. Issues Ergon. Sci. **20**(3), 223–249 (2019). https://doi.org/10.1080/1463922X.2017.1293187
83. Salvendy, G. (ed.): Handbook of Human Factors and Ergonomics: Salvendy/Handbook of Human Factors 4e. Wiley, Hoboken, NJ (2012). https://doi.org/10.1002/9781118131350
84. Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I., Reynolds, N.: Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health Care Inf. **26**(1), 100081 (2019). https://doi.org/10.1136/bmjhci-2019-100081
85. Hernández-Orallo, José: Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artif. Intell. Rev. **48**, 397–447 (2017)
86. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.: A survey on concept drift adaptation. ACM Comput. Surv. **46**(4), 1–37 (2014). https://doi.org/10.1145/2523813
87. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G.: Learning under Concept Drift: A Review. IEEE Trans. Knowl. Data Eng. pp. 1–1 (2018). https://doi.org/10.1109/TKDE.2018.2876857
88. Melman, T., de Winter, J.C.F., Abbink, D.A.: Does haptic steering guidance instigate speeding? A driving simulator study into causes and remedies. Accident Anal. Prevent. **98**, 372–387 (2017). https://doi.org/10.1016/j.aap.2016.10.016
89. Christian, G.: Partially automated driving as a fallback level of high automation. In: 6. Tagung Fahrerassistenzsysteme (2013)
90. Abbink, D.A., Carlson, T., Mulder, M., de Winter, J.C.F., Aminravan, F., Gibo, T.L., Boer, E.R.: A topology of shared control systems—finding common ground in diversity. IEEE Trans. Hum.-Mach. Syst. **48**(5), 509–525 (2018)
91. Abbink, D.A., Mulder, M., Boer, E.R.: Haptic shared control: smoothly shifting control authority? Cognit. Technol. Work **14**(1), 19–28 (2012). https://doi.org/10.1007/s10111-011-0192-5
92. Aristotle: Nicomachean Ethics, 2nd ed edn. Hackett Pub. Co, Indianapolis, Ind (1999)
93. Van de Poel, I.: Translating Values into Design Requirements. In: Michelfelder, D.P., McCarthy, N., Goldberg, D.E. (eds.) Philosophy and Engineering: Reflections on Practice, Principles and Process, pp. 253–266. Springer, New York (2013). https://doi.org/10.1007/978-94-007-7762-0_20
94. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)
95. Van de Poel, I., Royakkers, L.: Ethics, Technology, and Engineering: An Introduction. Wiley-Blackwell, Hoboken (2011)
96. Coeckelbergh, M.: Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci. Eng. Ethics (2019). https://doi.org/10.1007/s11948-019-00146-8
97. Douer, N., Meyer, J.: The responsibility quantification model of human interaction with automation. IEEE Trans. Autom. Sci. Eng. **17**(2), 1044–1060 (2020). https://doi.org/10.1109/TASE.2020.2965466
98. Latour, B., et al.: Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford (2005)
99. Johnson, J.: Mixing humans and nonhumans together: the sociology of a door-closer. Soc. Probl. **35**(3), 298–310 (1988)
100. Pendleton-Jullian, A.M., Brown, J.S.: Design Unbound: Designing for Emergence in a White Water World. MIT Press, Cambridge (2018)
101. Van der Bijl-Brouwer, M., Malcolm, B.: Systemic design principles in social innovation: a study of expert practices and design rationales. She Ji: J. Des. Econ. Innovat. **6**(3), 386–407 (2020). https://doi.org/10.1016/j.sheji.2020.06.001