MEANINGFUL HUMAN CONTROL OVER AI SYSTEMS: BEYOND TALKING THE TALK

A PREPRINT

Luciano Cavalcante Siebert[*,†], **Maria Luce Lupetti**[*], **Evgeni Aizenberg**[*], **Niek Beckers**[*], **Arkady Zgonnikov**[*],
Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M. Jonker,
Jeroen van den Hoven, Deborah Forster, Reginald L. Lagendijk
AiTech Interdisciplinary Research Program on Meaningful Human Control
Delft University of Technology
Delft, The Netherlands

ABSTRACT

How can humans remain in control of artificial intelligence (AI)-based systems designed to have autonomous capabilities? 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 (e.g., users, designers and developers, manufacturers, legislators). However, the relevant discussions around meaningful human control have so far not resulted in clear requirements for researchers, designers, and engineers. As a result, there is no consensus on how to assess whether a designed AI system is under meaningful human control, making the practical 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 four actionable properties which AI-based systems must have to be under meaningful human control. 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 these four properties are necessary for AI systems under meaningful human control, and provide possible directions to incorporate them into practice. We illustrate these properties with two use cases, automated vehicle and AI-based hiring. We believe these four properties will support practically-minded professionals to take concrete steps toward designing and engineering for AI systems that facilitate meaningful human control and responsibility.

Keywords Artificial intelligence · AI Ethics · Meaningful Human Control · Moral responsibility · Socio-technical systems

1 Introduction

Artificial intelligence (AI) algorithms have become widely accessible and robust enough to effectively support innovation in many services and sectors, such as manufacturing, healthcare, entertainment, education, banking, and infrastructures [96]. However, deploying such AI algorithms in human-inhabited environments also comes with the risk of inappropriate, undesirable, or unpredictable consequences [36]. The misinterpreted capabilities of AI, combined with their rapid impact in public and private spheres of life, calls for a careful alignment to moral values and societal norms [25, 26, 46, 85].

[*] Co-first authors.

[†] Correspondent author: L.CavalcanteSiebert@tudelft.nl

In this paper, we focus on one of the many concerns associated with AI: *moral responsibility*. How can designers, users, or other human agents be morally responsible for systems with autonomous capabilities to learn and adapt without direct human control? Similar to achieving autonomous capabilities, facilitating moral responsibility in human-AI systems (systems in which humans and AI algorithms interact) is a complex design issue that must be addressed early on [30]. The very autonomous capabilities that make AI algorithms useful complicate their assessment and predictability in the real-world as well as their validity over time. As a result, all systems based on AI, especially those with higher levels of autonomy, can and should be designed for appropriate human responsibility [75]. 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. Such situations represent so-called responsibility gaps [61]: circumstances with undesirable impacts where responsibility cannot be properly attributed to any person. Situations where responsibility gaps come to light include e.g., accidents involving highly-automated vehicles and the use of human-AI systems in recruitment processes where recruiters cannot control undesirable impacts such as discrimination and unfairness. Relevant humans (e.g., designers, developers, or users) should be able to act upon their moral responsibility in a manner that is equally complex and fine-grained as the socio-technical system to which the AI algorithm is applied.

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 [62]. 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, law, 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 decision-making capabilities. 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 [49, 68]. Second, the agent's decision-making process may be difficult to explain and predict, even for its programmer [33], 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 capabilities are overestimated by those interacting with it. As the system's design process may overlap with implementation and use [40], 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 capabilities 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 capabilities 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" [20], instead of regulated tests procedures, is urging academia and governments to take a stance in defining visions for trustworthy AI [46]. 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 capabilities [82]. A failure of an AI agent is not a "404 error page". It is a car accident, most likely with fatalities [50, 78]; it is an unfair and discriminatory distribution of wealth and services [53]; it is an unjust crime accusation based on ethnicity [4, 81]. We can only tackle this challenge if we acknowledge upfront that successful attribution and apportioning of responsibility should not be a matter of fortuitous allocation of praise or blame.

Developing concrete strategies to innovate in a responsible manner [89] is urgent and crucial to prevent such unintended and undesirable consequences, but also not to hinder the potential benefit that AI technologies can offer to society [36]. As a response, we see an increasingly growing effort from governments, the private sector, and academia to investigate their role and responsibility in developing a "good AI society" [19], which is leading to a proliferation of ethical guidelines and frameworks [42]. Although a convergence around five emerging ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy) [46] provides a shared understanding of the problems at stake, methods to properly mitigate associated risks are still in their infancy [36]. It is imperative that widely accepted methodologies and practices are established to bridge the gap between agreed upon ethical principles and the real-life behavior of human-AI systems, avoiding responsibility gaps.

The concept of *meaningful human control* [7, 8, 45, 75] was first proposed to address the problem of responsibility gaps in autonomous weapon systems, but is becoming a central concepts when discussing responsible AI [75].The core idea is that humans should ultimately remain in control of, and thus morally responsible for, the behavior of human-AI systems [1]. Nevertheless, meaningful human control has also received the critique to be an ill-defined concept [27] that ignores operational context [32] and does not provide concrete design guidelines [62].

This article aims to contribute to closing the gap between the theory of meaningful human control, as proposed by [75], 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 (Section 2). We then present a set of four properties that indicate whether and to what extent a human-AI system is under meaningful human control, discussing concrete methods and tools that can support addressing each property and illustrating them with respect to two case studies: automated vehicles and AI-based hiring (Section 3). Finally, we discuss the systemic and socio-technical nature of these properties and the need for multidisciplinary actions (Section 4) and conclude the paper (Section 5).

2 Related work on meaningful human control

The concept of meaningful human control was coined in the debates on autonomous weapon systems in order to formulate the conditions under which the pursuit and deployment of lethal autonomous weapons could be considered to be morally acceptable. While there is an overall consensus concerning the need for some form of human control over autonomous weapons, there are divergent and often conflicting views about what makes human control meaningful [7, 8, 32].

This discussion is no longer exclusive to the military domain. Meaningful human control is increasingly relevant and necessary as AI systems become more ubiquitous and autonomous, especially in non-forgiving scenarios in which fundamental human rights are at stake. The concept has already been applied to automated vehicles [10, 16, 62], including truck platooning [18], surgical robots [34], smart home systems [84], medical diagnosis [13], and content moderation in social media [94].

In the quest to prevent responsibility gaps, the concept of meaningful human control relies on the assumption that responsibility should always be attributed to human agents (e.g., users, operators, designers), not to AI agents. For this, we follow Santoni and Van den Hoven's [75] philosophical account towards two necessary conditions for having a human-AI system under meaningful human control (not precluding other necessary conditions):

(1) **Tracking** condition: in order 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.

The tracking and tracing conditions provide a valuable philosophical framework for meaningful human control. 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.

Building on [75], researchers developed frameworks to discuss and quantify factors affecting meaningful human control, especially for automated vehicles [43, 17, 62]. However, a range of key questions regarding the operationalization of tracking and tracing conditions to support the design and development of human-AI systems that can remain under meaningful human control remain unanswered. Whose and which moral reasons should account for the system's behavior? How are these reasons represented and operationalized? Who has influence and understanding and who is in control? Who is responsible for the system's (un)ethical behavior?

[1] 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, 75, 73]. As [62] 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."

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

The tracking and tracing conditions [75] provide a starting point towards thinking about what to consider when developing human-AI systems under meaningful human control. However, it is not trivial to translate these philosophical concepts towards more concrete design and engineering directions. We unpack the tracking and tracing conditions in four *actionable properties* which can more easily support the design, development, and evaluation of human-AI systems.

First, for the human-AI system 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. When designing for the tracking condition, it does not suffice to specify the *technical* conditions in which the system is designed to operate, i.e. the *operational design domain* (ODD). Designers should consider a larger design domain that captures which *values and societal norms* should be considered and respected during its design and operation: the *moral* ODD. Hence, we present as the first property that *a human-AI system should have an explicitly specified moral operational design domain (moral ODD), and the AI agent should adhere to the boundaries of this domain*.

Note that this moral ODD is not static: unforeseen and possibly detrimental behaviors can emerge during deployment of the system that might not be anticipated in the original moral ODD. Over time and across contexts, there might be developments in who the relevant stakeholders are, or what human reasons to include. Therefore, it is essential to consider the dynamic nature of human-AI systems in identifying the relevant humans, their reasons, and circumstances already during the design phase. Even with the best intentions and anticipations, unintended consequences are bound to emerge. This necessitates that we ensure that humans are in control of the system and are able to adjust the systems to their preferences. To answer to these concerns, we propose the second and third properties, inspired by the perception-action approach to complex human-machine interactions [28].

From a perception point-of-view, to decide what action to take and to subsequently perform that action, the relevant agents (human and AI) should be able to build appropriate and dynamic representations of each other and of the moral ODD. These representations include the agents' reasons, tasks, desired outcomes, role distributions, and preferences. We capture this in the second property, which requires that the *human and AI agents have appropriate and mutually compatible representations of the human-AI system*.

Considering the "action challenge", humans (e.g., users, developers) should be able to change the system's goals and behavior in order to track changing human reasons. In addition, humans should be able to intervene and correct the behavior of the system when unforeseen situations occur. 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. This supports humans on being aware of their responsibility and acting accordingly. We, therefore, define a third property positing that *relevant agents have ability and authority to control the system so that humans can act upon their responsibility*.

These first three properties contribute to the tracking condition of meaningful human control, as well as to the tracing condition, which requires at least one human who designs or interacts with the system to have a proper moral understanding of the system's effects in the world. However, the tracing condition also requires that the system's actions are traceable to the moral understanding of this human or humans. This tracing process should be transparent, explainable, and developed in ways that allows it to be inspectable. Furthermore, we argue that moral understanding of the system's effects should be demonstrated by, at least, those 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 [3], 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. This leads to the fourth and final property: *actions of the AI agents are explicitly linked to actions of humans who are aware of their moral responsibility*.

To summarize, we argue that a human-AI system under meaningful human control should possess each of the following four properties:

- **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.

Similar to the tracking and tracing conditions [75], these properties are necessary but not sufficient for a system to be under meaningful human control: while a system possessing all these properties may still not be completely under meaningful human control, missing one of these properties means that the human-AI system is surely not under meaningful human control. Moreover, these properties are not binary and are hard to quantify. Improving the system according to one or more of these properties will lead to better tracking or tracing, and therefore, more meaningful human control over that system, however defining "how much of these properties" is sufficient in a given context might often only be possible through a qualitative and situated analysis.

Furthermore, these properties in themselves do not immediately translate to concrete design guidelines. In the coming sections we will expand each of the properties and establish explicit links to conceptual frameworks and methodologies across the design and engineering domains to support practically-minded professionals. However, specific design approaches, evaluation methods and metrics, algorithms, and methodologies needed to implement these four properties are context- and system-specific. Still, to illustrate the usefulness of these four properties, we will discuss them with respect to two relevant cases of human-AI systems: automated vehicles and AI-based hiring. While both cases manifest an urgent need for meaningful human control in non-forgiving scenarios that strongly impact people's life (e.g., bodily harm, unfair decisions, discrimination), their differences with respect to time-constraints, embodiment, and involved stakeholders interestingly juxtapose different aspects of realizing these properties in human-AI systems.

- **Automated vehicle:** 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 capabilities 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.
- **AI-based hiring:** Job candidates applying for a vacancy go through a 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.

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

Setting explicit operational boundaries for a system is a complex process that asks for building rich knowledge about the operating conditions within which a system may operate, and selecting the acceptable ones. In this regard, we relate to and expand the concept of operational design domain (ODD) which 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 [71]. ODD specifications 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) [29]. However, the ODD is not just a mere description of a physical domain. Instead, it is a critical analysis and selection of scenarios that can be safely managed [52] and in which undesired consequences are minimized [29]. We believe it is a valuable concept not only for automated vehicles, but for human-AI systems in general.

The focus of the current concept of ODD is, however, still on the technical aspects of operation. In this perspective, technical development is often focused on expanding the boundaries of the ODD, but consideration of the wider societal implications is lacking. Similar to [14], we argue that the concept of ODD should also emphasize the broader social and ethical implications. We refer to this extended concept of ODD as a *moral operational design domain (moral ODD)*. We propose the moral ODD (Figure 1) as means to determine not only the domain in which the AI agent *can* operate from a purely functional perspective, but also the domain in which it *ought* or *should not* operate from a moral perspective.

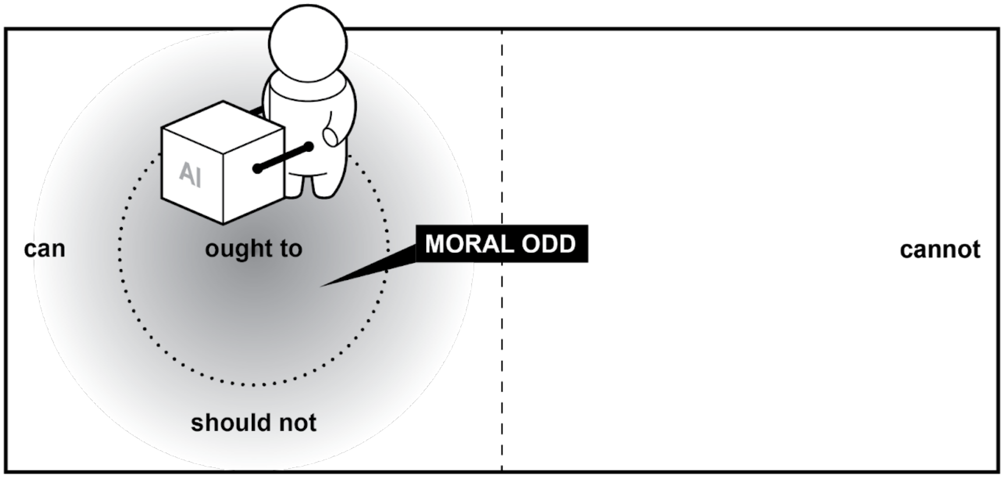

Figure 1: Property 1: Moral ODD. The human-AI system should operate within the boundaries of what it can do (for both the human and the AI agent) and within the moral boundaries of what it ought to do, i.e. the human-AI system should act according to the relevant moral reasons of the relevant stakeholders.

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 where complex human-AI systems are involved, 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 autonomy over self-representation is 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 [12, 15]. 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 [23, 38, 21], participatory approaches such as interviews, interactive workshops, and value-oriented coding of qualitative responses [37], which can be supported by natural language processing algorithms [57].

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 [67] and geofencing [59]. 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 [97]. 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 [69].

Moral constraints are arguably the most challenging to enforce. One possible way of imposing them is to set probabilistic guarantees on system outcomes [83]. 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 [14]. Hence, it is crucial that humans, not AI agents, are empowered to be aware of their responsibilities in order 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.

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 [48, 51, 95].

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 [51], 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 as overreliance [56]. Furthermore — crucial to achieve meaningful human control — these shared representations should include the human reasons identified in the moral ODD (Figure 2). As the elements of the shared representations can be time- and context-dependent, 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 [35].

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

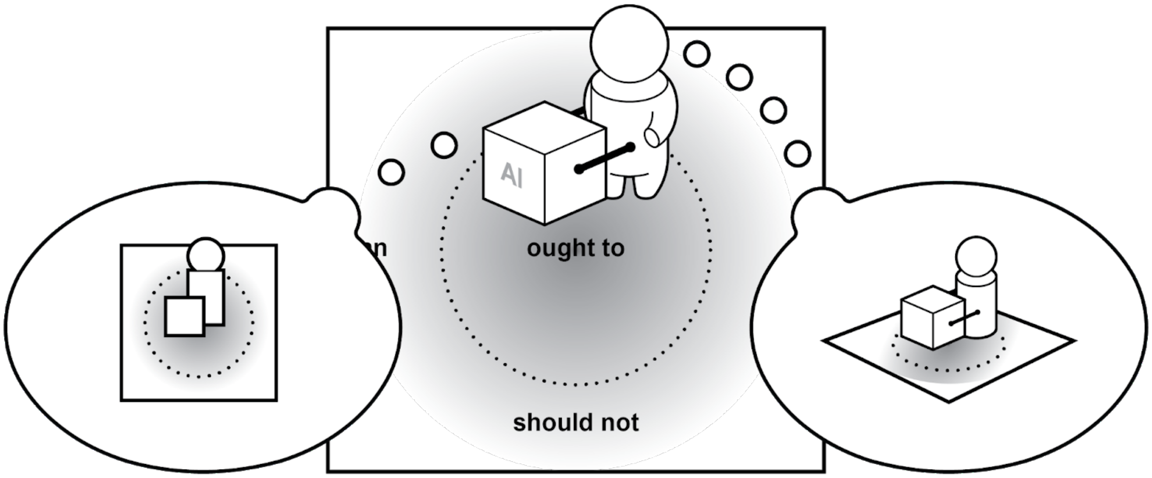

Figure 2: Property 2: The human and AI agents have appropriate and mutually compatible representations of the human-AI system and of each other's abilities and boundaries.

is co-active design [48]. 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 [51].

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 [3]), 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 [93]. 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 it's intentions before committing to a decision [70].

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 [77, 79], or by augmenting bottom-up, machine-learned representations with top-down symbolic representations [86, 60]. An alternative approach, value alignment, 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) [41]. Although IRL is likely not sufficient to infer human preferences from observed behaviour since human planning systematically deviates from the assumed global rationality [6], such approaches could still support agents to maintain aligned shared representations [66].

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 (Figure 3). For this, they need to have appropriate ability and authority to steer the behavior of the human-AI system, including overruling the AI agent through intervening and correcting behavior, setting new goals, or delegating sub-tasks. Furthermore, such actions need to be aligned with the abilities and control authority of the other agents in the system.

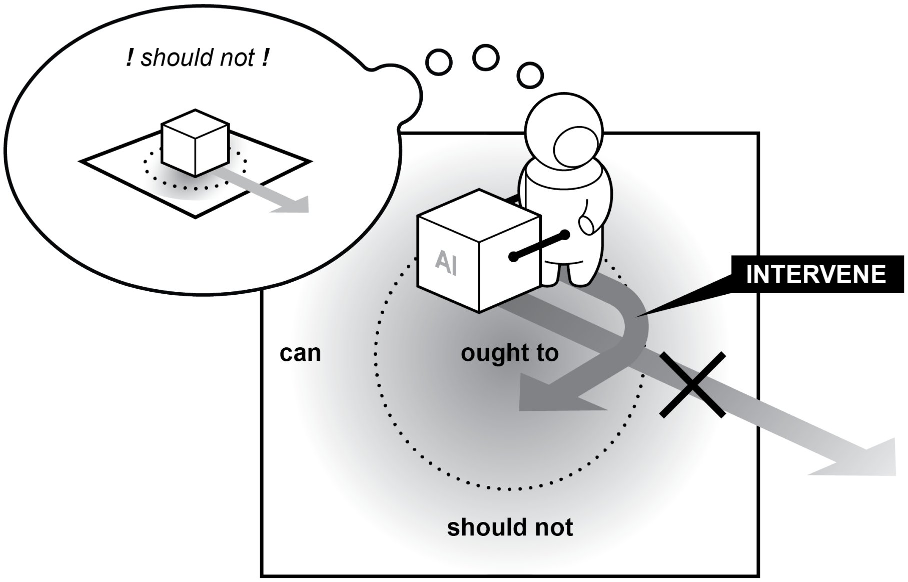

Figure 3: Property 3: The relevant humans and AI agents have the ability and authority to control the system so that humans can act upon their responsibility, e.g., if the human recognizes that a given situation might bring the system outside the moral ODD, they can intervene to avoid this.

Flemisch et al. [35] 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 [64]. 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 [74, 76], under the umbrella of Value Sensitive Design [37], 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 [92, 91].

The second component of the account proposed in [35] 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 [35, 43, 54]. The same might happen for a manager working with an AI-based hiring system if the manager, who has the final authority over who to hire, merely signs off on the suggestions of the AI agent, without substantive engagement in the assessment process.

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 [72] 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 [80, 54]. 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 [44], and ethology (related to animal behavior) [68]. Designing for appropriate authority and ability also requires us to expand the scope of design from human-AI interactions to social and organizational practices [75]. 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 [39, 58]. 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 [63], 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 [22] 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 [1]. 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 [2]. 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 and inspectable link(s) between actions of the system and corresponding human morally-loaded decisions and actions. We acknowledge that such inspectable link(s) might be a more demanding form of tracing than what was originally proposed in [75], nevertheless we deem it necessary to enable the tracing condition to be inspectable. 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 (Figure 4):

(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.

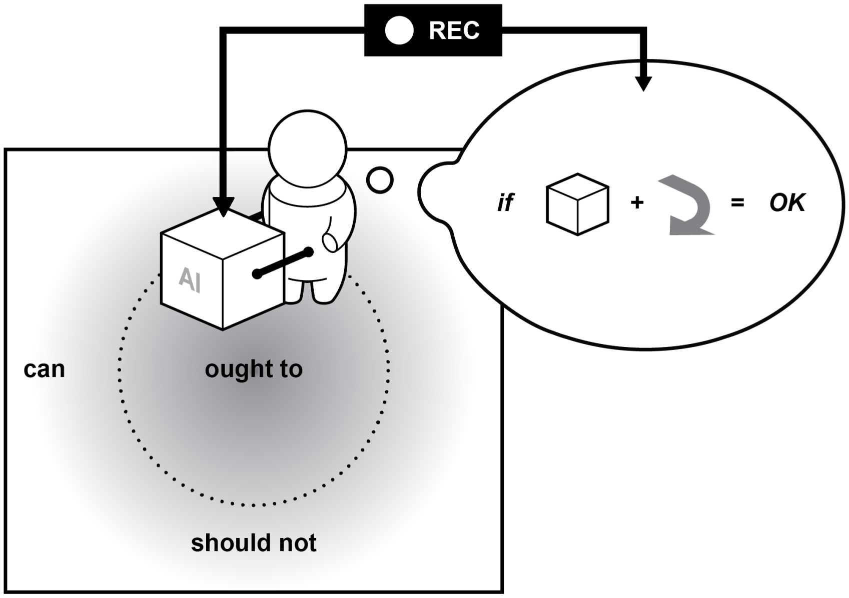

Figure 4: Property 4: Actions of the AI agents are explicitly linked to actions of humans who are aware of their moral responsibility. The forward link starts on the human and indicates that whenever a human makes a decision with moral implications that affects the system's behavior, that human should 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).

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 [5, 25]. 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 [87] 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 [9]. 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 [62]. 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 could be used to remind hiring managers of their duty to reduce discrimination in the hiring process.

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 [88]. 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 [24]. 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 [31] 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) [55] 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", [47] 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.).

4 Discussion

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 to implement them in practice. These properties unpack the tracking and tracing conditions of meaningful human control [75] 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 [29], ontological modeling [15], co-active design [48], shared mental models [51], shared control [1], value alignment [66], and consistency of ability, control authority, and responsibility [35]. In addition to establishing explicit links between the concept of meaningful human control and these frameworks, 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 the 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 necessary 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. It is not reasonable to expect that design and regulation would 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 [65], where the focus shifts to designing the social, physical, and technical infrastructures that 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. In fact, 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 [75]. 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 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 [90] 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

The need for meaningful human control should not be limited to autonomous weapon 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. In this paper, we contribute to this debate by identifying a set of four properties aimed to support practice-minded professionals to design and develop systems that can incorporate a meaningful form of human control. With these four properties we have realized two goals: (1) contributed to closing the gap between the theory and practice of 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. 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.

Acknowledgement

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

This work was supported by the AiTech program of the Delft University of Technology.

References

[1] Abbink, D.A., Carlson, T., Mulder, M., de Winter, J.C.F., Aminravan, F., Gibo, T.L., Boer, E.R., 2018. A Topology of Shared Control Systems—Finding Common Ground in Diversity. IEEE Transactions on Human-Machine Systems 48, 509–525. doi:doi:10/gfbkmk.

[2] Abbink, D.A., Mulder, M., Boer, E.R., 2012. Haptic shared control: smoothly shifting control authority? Cognition, Technology & Work 14, 19–28. doi:doi:10.1007/s10111-011-0192-5.

[3] Aler Tubella, A., Theodorou, A., Dignum, F., Dignum, V., 2019. Governance by Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Macao, China. pp. 5787–5793. doi:doi:10.24963/ijcai.2019/802.

[4] Angwin, J., Jeff Larson, Surya Mattu, Lauren Kirchner, 2016. Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica .

[5] Aristotle, 1999. Nicomachean ethics. 2nd ed ed., Hackett Pub. Co, Indianapolis, Ind.

[6] Armstrong, S., Mindermann, S., 2018. 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, Curran Associates, Inc. URL: `https://proceedings.neurips.cc/paper/2018/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf`.

[7] Article 36, 2014. 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).

[8] Article 36, 2015. Killing by Machine: Key Issues for Understanding Meaningful Human Control.

[9] 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., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115. doi:doi:10/ggqs5w.

[10] 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., 2019. 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, SPIE, Strasbourg, France. p. 11. doi:doi:10.1117/12.2533740.

[11] Behymer, K.J., Flach, J.M., 2016. From Autonomous Systems to Sociotechnical Systems: Designing Effective Collaborations. She Ji: The Journal of Design, Economics, and Innovation 2, 105–114. doi:doi:10/ggsqcq.

[12] Bettini, C., Brdiczka, O., Henricksen, K., Indulska, J., Nicklas, D., Ranganathan, A., Riboni, D., 2010. A survey of context modelling and reasoning techniques. Pervasive and mobile computing 6, 161–180. doi:doi:10.1016/j.pmcj.2009.06.002.

[13] Braun, M., Hummel, P., Beck, S., Dabrock, P., 2020. Primer on an ethics of AI-based decision support systems in the clinic. Journal of Medical Ethics , medethics–2019–105860doi:doi:10.1136/medethics-2019-105860.

[14] Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., Porter, Z., 2020. Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. Artificial Intelligence 279, 103201. doi:doi:10.1016/j.artint.2019.103201.

[15] Cabrera, O., Franch, X., Marco, J., 2019. 3LConOnt: a three-level ontology for context modelling in context-aware computing. Software & Systems Modeling 18, 1345–1378. doi:doi:10.1007/s10270-017-0611-z.

[16] Calvert, S.C., Heikoop, D.D., Mecacci, G., Van Arem, B., 2019. A human centric framework for the analysis of automated driving systems based on meaningful human control. Theoretical Issues in Ergonomics Science , 1–29doi:doi:10.1080/1463922X.2019.1697390.

[17] Calvert, S.C., Mecacci, G., 2020. 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 , 1–1doi:doi:10.1109/OJITS.2020.3021461.

[18] Calvert, S.C., Mecacci, G., Heikoop, D.D., de Sio, F.S., 2018. Full platoon control in Truck Platooning: A Meaningful Human Control perspective, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, Maui, HI. pp. 3320–3326. doi:doi:10.1109/ITSC.2018.8570013.

[19] Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., Floridi, L., 2018. Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. Science and Engineering Ethics doi:doi:10.1007/s11948-017-9901-7.

[20] Chen, L., Wilson, C., 2017. Observing algorithmic marketplaces in-the-wild. ACM SIGecom Exchanges 15, 34–39. doi:doi:10.1145/3055589.3055594.

[21] 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., 2002. Public health ethics: mapping the terrain. The Journal of Law, Medicine & Ethics 30, 170–178. doi:doi:10.1111/j.1748-720X.2002.tb00384.x.

[22] Christian, G., 2013. Partially automated driving as a fallback level of high automation, in: 6. Tagung Fahrerassistenzsysteme.

[23] Coeckelbergh, M., 2013. Drones, information technology, and distance: mapping the moral epistemology of remote fighting. Ethics and information technology 15, 87–98. doi:doi:10.1007/s10676-013-9313-6.

[24] Coeckelbergh, M., 2019. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. Science and Engineering Ethics doi:doi:10.1007/s11948-019-00146-8.

[25] Coeckelbergh, M., 2020. AI ethics. The MIT press essential knowledge series, The MIT Press, Cambridge, MA.

[26] Cruz, J., 2019. Shared Moral Foundations of Embodied Artificial Intelligence, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 139–146.

[27] Cummings, M., 2019. Lethal Autonomous Weapons: Meaningful Human Control or Meaningful Human Certification? IEEE Technology and Society Magazine 38, 20–26. doi:doi:10.1109/MTS.2019.2948438.

[28] Cutsuridis, V., Hussain, A., Taylor, J.G. (Eds.), 2011. Perception-Action Cycle. Springer New York, New York, NY. doi:doi:10.1007/978-1-4419-1452-1.

[29] Czarnecki, K., 2018. Operational Design Domain for Automated Driving Systems - Taxonomy of Basic Terms doi:doi:10/ggr6kg.

[30] 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., 2018. Ethics by Design: Necessity or Curse?, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, ACM, New Orleans LA USA. pp. 60–66. doi:doi:10.1145/3278721.3278745.

[31] Douer, N., Meyer, J., 2020. The Responsibility Quantification Model of Human Interaction With Automation. IEEE Transactions on Automation Science and Engineering 17, 1044–1060. doi:doi:10.1109/TASE.2020.2965466.

[32] Ekelhof, M., 2019. Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation. Global Policy 10, 343–348. doi:doi:10.1111/1758-5899.12665.

[33] European Parliamentary Research Service (EPRS), 2020. The ethics of artificial intelligence: Issues and initiatives. Panel for the Future of Science and Technology PE 634.452.

[34] Ficuciello, F., Tamburrini, G., Arezzo, A., Villani, L., Siciliano, B., 2019. Autonomy in surgical robots and its meaningful human control. Paladyn, Journal of Behavioral Robotics 10, 30–43. doi:doi:10/ggqkzn.

[35] Flemisch, F., Heesen, M., Hesse, T., Kelsch, J., Schieben, A., Beller, J., 2012. Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations. Cognition, Technology & Work 14, 3–18. doi:doi:10/c7xmbh.

[36] Floridi, L., Cowls, J., King, T.C., Taddeo, M., 2020. How to Design AI for Social Good: Seven Essential Factors. Science and Engineering Ethics 26, 1771–1796. doi:doi:10.1007/s11948-020-00213-5.

[37] Friedman, B., Hendry, D.G., 2019. Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press.

[38] Galliott, J., 2015. Military robots: Mapping the moral landscape. Ashgate Publishing, Ltd.

[39] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, A., 2014. A survey on concept drift adaptation. ACM Computing Surveys 46, 1–37. doi:doi:10.1145/2523813.

[40] Giaccardi, E., Redström, J., 2020. Technology And More-Than-Human Design. Design Issue 36.

[41] Hadfield-Menell, D., Russell, S.J., Abbeel, P., Dragan, A., 2016. Cooperative Inverse Reinforcement Learning. Advances in Neural Information Processing Systems 29.

[42] Hagendorff, T., 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines 30, 99–120. doi:doi:10.1007/s11023-020-09517-8.

[43] Heikoop, D.D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., Van Arem, B., 2019. Human behaviour with automated driving systems: a quantitative framework for meaningful human control. Theoretical Issues in Ergonomics Science 20, 711–730. doi:doi:10.1080/1463922X.2019.1574931.

[44] Hernández-Orallo, José, 2017. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artificial Intelligence Review 48, 397–447.

[45] Horowitz, M.C., Scharre, P., 2015. Meaningful Human Control in Weapon Systems: A Primer. working paper , 15.

[46] Jobin, A., Ienca, M., Vayena, E., 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence 1, 389–399. doi:doi:10.1038/s42256-019-0088-2.

[47] Johnson, J., 1988. Mixing humans and nonhumans together: The sociology of a door-closer. Social problems 35, 298–310.

[48] Johnson, M., Bradshaw, J.M., Feltovich, P.J., Jonker, C.M., Van Riemsdijk, M.B., Sierhuis, M., 2014. Coactive Design: Designing Support for Interdependence in Joint Activity. Journal of Human-Robot Interaction 3, 43. doi:doi:10.5898/JHRI.3.1.Johnson.

[49] Johnson, N., Zhao, G., Hunsader, E., Qi, H., Johnson, N., Meng, J., Tivnan, B., 2013. Abrupt rise of new machine ecology beyond human response time. Scientific Reports 3, 2627. doi:doi:10.1038/srep02627.

[50] Johnston, P., Harris, R., 2019. The Boeing 737 MAX saga: lessons for software organizations. Software Quality Professional 21, 4–12.

[51] Jonker, C.M., Van Riemsdijk, M.B., Vermeulen, B., 2011. Shared Mental Models, in: De Vos, M., Fornara, N., Pitt, J.V., Vouros, G. (Eds.), Coordination, Organizations, Institutions, and Norms in Agent Systems VI, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 132–151.

[52] Koopman, P., Fratrik, F., 2019. How Many Operational Design Domains, Objects, and Events? , 4.

[53] Korinek, A., Stiglitz, J.E., 2017. Artificial intelligence and its implications for income distribution and unemployment. Technical Report 0898-2937. National Bureau of Economic Research.

[54] 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., 2019. A human factors perspective on automated driving. Theoretical Issues in Ergonomics Science 20, 223–249. doi:doi:10.1080/1463922X.2017.1293187.

[55] Latour, B., et al., 2005. Reassembling the social: An introduction to actor-network-theory. Oxford university press.

[56] Lee, J.D., See, K.A., 2004. Trust in Automation: Designing for Appropriate Reliance. Human Factors , 31.

[57] Liscio, E., Van der Meer, M., Siebert, L.C., Jonker, C.M., Mouter, N., Murukannaiah, P.K., 2021. Axies: Identifying and evaluating context-specific values, in: Proceedings of the 20th international conference on autonomous agents and MultiAgent systems, pp. 799–808.

[58] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G., 2018. Learning under Concept Drift: A Review. IEEE Transactions on Knowledge and Data Engineering , 1–1doi:doi:10.1109/TKDE.2018.2876857.

[59] Maiouak, M., Taleb, T., 2019. Dynamic Maps for Automated Driving and UAV Geofencing. IEEE Wireless Communications 26, 54–59. doi:doi:10.1109/MWC.2019.1800544.

[60] Marcus, G., 2020. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. arXiv:2002.06177 [cs] arXiv:2002.06177.

[61] Matthias, A., 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and information technology 6, 175–183. doi:doi:10.1007/s10676-004-3422-1.

[62] Mecacci, G., Santoni de Sio, F., 2020. Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. Ethics and Information Technology 22, 103–115. doi:doi:10.1007/s10676-019-09519-w.

[63] Melman, T., de Winter, J., Abbink, D., 2017. Does haptic steering guidance instigate speeding? a driving simulator study into causes and remedies. Accident Analysis & Prevention 98, 372–387. doi:doi:10.1016/j.aap.2016.10.016.

[64] Parasuraman, R., Sheridan, T., Wickens, C., 2000. A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 30, 286–297. doi:doi:10/c6zf92.

[65] Pendleton-Jullian, A.M., Brown, J.S., 2018. Design Unbound: Designing for Emergence in a White Water World. MIT Press.

[66] Peysakhovich, A., 2019. 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, Association for Computing Machinery, New York, NY, USA. pp. 409–415. doi:doi:10.1145/3306618.3314259.

[67] Primatesta, S., Scanavino, M., Guglieri, G., Rizzo, A., 2020. 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), IEEE, Athens, Greece. pp. 641–650. doi:doi:10.1109/ICUAS48674.2020.9213982.

[68] 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., 2019. Machine behaviour. Nature 568, 477–486. doi:doi:10/gfzvhx.

[69] Robbins, S., 2020. AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines. AI & SOCIETY 35, 391–400. doi:doi:10.1007/s00146-019-00891-1.

[70] Sadigh, D., Landolfi, N., Sastry, S.S., Seshia, S.A., Dragan, A.D., 2018. Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state. Autonomous Robots 42, 1405–1426. doi:doi:10.1007/s10514-018-9746-1.

[71] SAE, 2018. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Technical Report. SAE International. doi:doi:10.4271/J3016_201806.

[72] Salvendy, G. (Ed.), 2012. Handbook of Human Factors and Ergonomics: Salvendy/Handbook of Human Factors 4e. John Wiley & Sons, Inc., Hoboken, NJ, USA. doi:doi:10.1002/9781118131350.

[73] Santoni de Sio, F., Mecacci, G., 2021. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philosophy & Technology , 1–28.

[74] Santoni de Sio, F., Robichaud, P., Vincent, N.A., 2014. Who Should Enhance? conceptual and Normative Dimensions of Cognitive Enhancement. HUMANA.MENTE Journal of Philosophical Studies 7, 179–197.

[75] Santoni de Sio, F., Van den Hoven, J., 2018. Meaningful Human Control over Autonomous Systems: A Philosophical Account. Frontiers in Robotics and AI 5, 15. doi:doi:10/gf597h.

[76] Santoni de Sio, F., Van Wynsberghe, A., 2016. When Should We Use Care Robots? the Nature-of-Activities Approach. Science and Engineering Ethics 22, 1745–1760. doi:doi:10/ggp7nt.

[77] Schürmann, T., Beckerle, P., 2020. Personalizing Human-Agent Interaction Through Cognitive Models. Frontiers in Psychology 11, 8. doi:doi:10.3389/fpsyg.2020.561510.

[78] Serter, B., Beul, C., Lang, M., Schmidt, W., 2017. Foreseeable Misuse in Automated Driving Vehicles-The Human Factor in Fatal Accidents of Complex Automation. Technical Report 0148-7191. SAE Technical Paper.

[79] Siebinga, O., Zgonnikov, A., Abbink, D., 2021. Validating human driver models for interaction-aware automated vehicle controllers: A human factors approach. arXiv:2109.13077 [cs] `arXiv:2109.13077`.

[80] Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I., Reynolds, N., 2019. Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health & Care Informatics 26, e100081. doi:doi:10.1136/bmjhci-2019-100081.

[81] Sweeney, L., 2013. Discrimination in online ad delivery. Queue 11, 10–29. doi:doi:10.1145/2460276.2460278.

[82] Taplin, J., 2017. Move fast and break things: How Facebook, Google, and Amazon have cornered culture and what it means for all of us. Pan Macmillan.

[83] Thomas, P.S., da Silva, B.C., Barto, A.G., Giguere, S., Brun, Y., Brunskill, E., 2019. Preventing undesirable behavior of intelligent machines. Science 366, 999–1004. doi:doi:10/ggd3zx.

[84] Umbrello, S., 2020. Meaningful Human Control over Smart Home Systems: A Value Sensitive Design Approach. Humana.Mente Journal of Philosophical Studies 12.

[85] Umbrello, S., De Bellis, A.F., 2018. A value-sensitive design approach to intelligent agents. Artificial Intelligence Safety and Security (2018) CRC Press (. ed) Roman Yampolskiy .

[86] Van Bekkum, M., de Boer, M., Van Harmelen, F., Meyer-Vitali, A., ten Teije, A., 2021. Modular Design Patterns for Hybrid Learning and Reasoning Systems: a taxonomy, patterns and use cases. arXiv:2102.11965 [cs] `arXiv:2102.11965`.

[87] Van de Poel, I., 2013. Translating Values into Design Requirements, in: Michelfelder, D.P., McCarthy, N., Goldberg, D.E. (Eds.), Philosophy and Engineering: Reflections on Practice, Principles and Process. Springer, pp. 253–266. doi:doi:10.1007/978-94-007-7762-0_20.

[88] Van de Poel, I., Royakkers, L., 2011. Ethics, Technology, and Engineering: An Introduction.

[89] Van den Hoven, J., 2013. Value Sensitive Design and Responsible Innovation, in: Owen, R., Bessant, J., Heintz, M. (Eds.), Responsible Innovation. John Wiley & Sons, Ltd, Chichester, UK, pp. 75–83. doi:doi:10.1002/9781118551424.ch4.

[90] Van der Bijl-Brouwer, M., Malcolm, B., 2020. Systemic Design Principles in Social Innovation: A Study of Expert Practices and Design Rationales. She Ji: The Journal of Design, Economics, and Innovation 6, 386–407. doi:doi:10.1016/j.sheji.2020.06.001.

[91] Van der Waa, J., Van Diggelen, J., Cavalcante Siebert, L., Neerincx, M., Jonker, C., 2020. 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. Springer International Publishing, Cham. volume 12187, pp. 203–220. doi:doi:10.1007/978-3-030-49183-3_16.

[92] Van Diggelen, J., Johnson, M., 2019. Team Design Patterns, in: Proceedings of the 7th International Conference on Human-Agent Interaction, ACM, Kyoto Japan. pp. 118–126. doi:doi:10.1145/3349537.3351892.

[93] Vicente, K., Rasmussen, J., 1992. Ecological interface design: theoretical foundations. IEEE Transactions on Systems, Man, and Cybernetics 22, 589–606. doi:doi:10.1109/21.156574.

[94] Wagner, B., 2019. Liable, but Not in Control? ensuring Meaningful Human Agency in Automated Decision-Making Systems. Policy & Internet 11, 104–122. doi:doi:10/gfv5sz.

[95] Wilson, J.R., Rutherford, A., 1989. Mental models: Theory and application in human factors. Human Factors 31, 617–634. doi:doi:10/gf4xtf.

[96] WIPO, 2019. WIPO Technology Trends 2019: Artificial Intelligence.

[97] Yavrucuk, I., Prasad, J.V.R., Unnikrishnan, S., 2009. Envelope Protection for Autonomous Unmanned Aerial Vehicles. Journal of Guidance, Control, and Dynamics 32, 248–261. doi:doi:10.2514/1.35265.