Meaningful Human Control and Responsibility Gaps in AI: No Culpability Gap, but Accountability and Active Responsibility Gap

Tatdanai Khomkhunsorn[✉]
Chulalongkorn University

(Received: 1st December 2024; Revised: 31st March 2025; Accepted: 3rd April 2025)

Abstract

At the current stage of technological development, the rapid advancement of Artificial Intelligence (AI) has given rise to various ethical concerns. Among these, the "Responsibility Gap" notion has appeared as a prominent issue. Within the scholarly literature, ethicists primarily focus on culpability (or blameworthiness). The central question is: when the development or use of AI results in morally harmful outcomes, who bears moral responsibility? This article argues that moral responsibility encompasses multiple distinct forms, each fulfilling specific functions within a society, especially in the context of AI development and application. Then, three forms of responsibility are considered: culpability, accountability, and active responsibility. Each carries unique social and ethical implications. Drawing on the concept of "meaningful human control," which serves as a foundational framework, this article contends that the gap in culpability is not as significant or troubling as often suggested in existing research. Instead, the more pressing ethical challenges are associated with gaps in accountability and active responsibility. To address these challenges, this article elaborates on the "tracing condition," a key element of meaningful human control, to mitigate and prevent morally harmful outcomes and the absence of human responsibility in the age of AI.

Keywords: meaningful human control, culpability, accountability, active responsibility, responsible AI

[✉] tatdanai5708@gmail.com

Introduction

Matthias (2004) first introduced this issue in his seminal article "The Responsibility Gap." He highlighted the distinctive properties of AI systems, referred to as "learning automata" (or machine learning), which possess the ability to autonomously learn, process information, and make decisions on their own through interactions with their environment. These systems often operate with minimal human prediction or control. However, normative practices of moral responsibility ascription are typically grounded in the principle of control and foresight, enabling individuals to anticipate and influence the consequences of their actions (Fischer & Ravizza, 1998). However, the emergence of autonomous AI systems has disrupted this framework due to the adaptive capabilities of the learning system, which can act without human intervention, complicating the assignment of responsibility when undesirable outcomes occur.

Since Matthias's initial work, many ethicists have engaged in the Responsibility Gap debate, either directly or indirectly responding to his claims (Tigard, 2020). As Santoni de Sio and Mecacci (2021) note, the Responsibility Gap has become one of the most widely discussed topics in AI ethics. Broadly, the debate can be divided into three perspectives: First, some scholars argue that the Responsibility Gap represents an inherent, unbridgeable, and insurmountable challenge (Matthias, 2004; Sparrow, 2007; Danaher, 2016). Thus, Matthias proclaimed that society faces a dilemma: either we continue to develop and use AI, thereby abandoning traditional practices of assigning responsibility, or we preserve human responsibility by ceasing the use and development of AI altogether.

Second, some authors contend that while the Responsibility Gap is real, it can be addressed through indirect ascription. For instance, responsibility can be ascribed to the humans involved in the AI's operation and development—such as designers, developers, engineers, and operators, which Goetze (2022) calls "vicarious responsibility." Then, the gap can be bridged. Alternatively, the notion of agency can be reconceptualized by treating human-AI interactions as "a unit" to which responsibility can be attributed (Hason, 2009; Nyholm, 2017; Piyanat, 2022). Third, some scholars deny the existence of a genuine responsibility gap, arguing that traditional frameworks for attributing responsibility, in the sense that responsibility lies in human agents involved, remain applicable even in the context of AI. For example, Kohler, Roughley, and Saurer assert that the advent of AI has not fundamentally altered established moral and legal paradigms (Köhler et al., 2017; Tigard, 2020; Hindriks & Veluwenkamp, 2022; Königs, 2022).

A recurring theme in the literature on AI ethics, as I suggested above, is the concern over moral culpability (blameworthiness)[1] in the context of AI use and development. While this is often the focus of debate, this article argues that moral responsibility encompasses multiple distinct forms, each serving specific societal functions. Here, I propose examining three forms of responsibility that are relevant—though not exclusive—to the development and application of AI: culpability, accountability, and active responsibility, each carrying unique ethical and social implications. Drawing on the concept of "meaningful human

[1] Some authors use the term *accountability* to denote the consequences that human agents must face when they commit wrongdoing, such as legal liability or blameworthiness ((Kohler et al., 2017; Konigs, 2022; Tigard, 2020A, 2020B), which in this sense is synonymous with the *culpability* I use here. However, the notion of accountability I adopt in this article is quite different from the above usage. Namely, the perception of an action is experienced as 'owned by' or attributable to human agents who identify themselves experientially with that action and then are obligated to explain and justify their decisions and actions (van de Poel & Sand, 2018; Santoni de Sio & Mecacci, 2021).

control" (MHC) developed by many scholars (Santoni de Sio and van den Hoven, 2018; Mecacci and Santoni de Sio, 2019; Cavalcante Siebert et al., 2022) and some other conceptual resources existing in the literature, I contend that the so-called culpability gap is often overstated in existing research and, ultimately, does not exist. Instead, the more pressing ethical challenges lie in gaps related to accountability and active responsibility.

The accountability gap arises when relevant agents fail to perceive the outcomes of AI operations as their own, which can be attributed to several factors. These include technical issues, such as the opacity of AI decision-making processes—commonly referred to as the "black box" problem— and, more importantly, which I am concerned with here, psychological factors, such as the experiential impacts of automation bias or over-reliance on AI from which "moral buffer, a form of psychological distancing, is created which allows people to ethically distance themselves from their actions." (Cummings, 2006, p. 3)

This lack of accountability, in turn, exacerbates an active responsibility gap characterized by a lack of awareness that others may hold legitimate moral reactions toward them for the consequences of their actions because of the roles they are in and the sort of relation they have with AI. From my standpoint, this is an urgent problem, given that it is essentially crucial for the agenda of responsible AI, which society as a whole should strive for.

To address these challenges, this article elaborates on the "tracing condition," a critical element of meaningful human control, as a mechanism to mitigate and prevent scenarios where no human involved feels responsible. Cultivating this sense of responsibility is a prerequisite for exercising accountability in the face of potentially harmful outcomes resulting from AI.

The article is structured as follows: First, I examine the problem of the alleged culpability gap in existing literature. Using the concept of MHC and other conceptual resources, I argue that culpability gaps are, at worst, less troubling than often assumed or, at best, non-existent. Second, I articulate the genuine ethical challenge of responsibility gaps, focusing on accountability and active responsibility. Finally, I elaborate on the tracing condition to address these responsibility gaps, followed by a conclusion.

2. Responsibility Gap as a Culpability Gap

2.1 The Meaning and Societal Function of Culpability

The notion of responsibility, as understood in normative contexts and reflected in the practice of assigning responsibility in human societies, can generally be categorized into two main types: *backward-looking* and *forward-looking*. The former often pertains to the evaluation of the blameworthiness of actions performed by individuals or collectives (e.g., organizations, political parties, or corporations) and is primarily guided by considerations of fairness. While the criteria for assessing blameworthiness may vary, three key elements, namely, intention, control, and knowledge, are commonly central to such evaluations, collectively constituting the notion of *culpability* (Santoni de Sio & Mecacci, 2021, p. 1061). By contrast, forward-looking responsibility involves obligations tied to an agent's role or duty to see to it that some desirable outcomes occur or to prevent harm, thereby fulfilling moral or societal expectations. (van de Poel & Sand, 2018)

This section focuses on culpability, as much of the recent literature on the responsibility gap—discussed in the previous section—has centered on this dimension of responsibility.

To clarify the concept of moral culpability, it is helpful to refer to an influential view in philosophy that bypasses the problem of free will and determinism. According to this view, when a person is morally responsible, they become a fitting target for reactive attitudes such as praise, blame, or punishment (Strawson, 1974). For instance, consider a person who intentionally shoots someone. If they exercised control over their actions, understood the consequences of pulling the trigger, and acted with intent to shoot, then it is fair for society to hold them culpable for the victim's death.

Let me contrast this with another scenario: imagine someone unknowingly consumes a drink spiked with an intoxicating drug (because someone secretly put it in) at a restaurant. When driving home, the drug impairs his motor skills, awareness, and sensitivity, causing him to lose control of the vehicle and fatally strike a pedestrian. In this case, the individual lacked control, was unaware of the risk, and had no intent to harm. Consequently, it would not be appropriate or fair to hold him culpable for the pedestrian's death[2] (even though he arguably should have a kind of "agent-regret"). In such a case, the individual is *excused* from culpability because the conditions needed for blame were not met.[3]

The primary societal function of culpability is its desert function, which satisfies the demands of retributive justice based on our moral intuition (Danaher, 2016; Doorn, 2009). In this context, culpability serves to apportion blame to moral agents in proportion to what they *deserve*, with fairness serving as the guiding principle in assigning responsibility. Furthermore, culpability also serves as a regulatory or incentive function. The practice of attributing responsibility creates an incentive structure that encourages moral agents to align their behavior with certain societal norms. This gives agents the motivational reasons to avoid blame and seek praise by complying with or refraining from violating these norms. This function ties backward-looking responsibility to forward-looking obligations (Doorn, 2009), such as those found in professional or ethical domains; for example, in engineering ethics, forward-looking responsibility requires agents to exercise reasonable care by anticipating and mitigating risks that might harm others or broader society. If agents fail to meet this standard—through negligence or recklessness—they can be deemed deserving of some degree of blame and held culpable even in the absence of intent, given that they have sufficient control and knowledge of the risks involved and prospective responsibility associated with the role obligation they should exercise.

As outlined earlier, this serves as a brief overview of the normative framework for assigning responsibility, along with its social and moral functions. In the following subsection,

[2] In a case where someone drinks alcohol themselves and causes a fatal accident while driving, even if there was no intention or control at the moment of the crash, they are still held responsible for the victim's life. This is because it is foreseeable that doing so carries a risk of harm to others, making it appropriate for them to be held culpable for the consequences.

[3] Additionally, in the case of young children, pets, individuals with mental health conditions, or even AI, they cannot be held morally responsible because there is widespread consensus that, although AI may be considered an intentional agent (Floridi & Sanders, 2004; Nyholm, 2017; Himmelreich, 2019), it does not possess moral agency on its own. Due to distinct agential conditions (Tigard, 2020B), these agents are exempt from normal responsibility ascription.

I will clarify the argument asserting the existence of a culpability gap stemming from the development of AI, with particular attention to the proposal put forward by Matthias.

2.2 The Problem of the Alleged Culpability Gap

Matthias was the first to explicitly argue that the advent of artificial intelligence (AI), particularly systems employing "machine learning," poses a fundamental threat to this established practice of responsibility attribution. He contends that this development creates what he terms a "responsibility gap." To elucidate his position, let us consider his words:

> Now it can be shown that there is an increasing class of machine actions, where the traditional ways of ascription of responsibility are not compatible with *our sense of justice and moral framework of society* [my italics] because nobody has enough control over the machine's actions to be able to assume the responsibility for them. These cases constitute what we will call the responsibility gap. (Matthias, 2004, p. 6–7)

The core of Matthias's argument is that the emergence of machine learning systems, such as self-driving cars or autonomous weapon systems, which are capable of autonomously determining their actions without direct human intervention, renders it inappropriate to assign responsibility to the operator or manufacturer in cases where morally harmful outcomes occur. This is because doing so would conflict with our "sense of justice and moral framework of society," which normally uses culpability as the primary criterion for responsibility attribution. As a result, a culpability gap arises. Even if the attribution of responsibility seems appropriate, there is no suitable target for ascribing moral responsibility.

To clarify, the culpability gap refers to situations in which a morally harmful outcome results from the autonomous actions of AI[4], in which blame is demanding, and yet no one can justly or reasonably be held responsible. On the one hand, AI systems themselves cannot bear moral responsibility because they lack the requisite mental capacities, such as moral competence or sensitivity to moral reasons. On the other hand, human agents cannot be held culpable either, as they lack sufficient knowledge and control over the AI's autonomous actions. Consequently, assigning responsibility to humans in such cases violates the principle of fairness that underpins principal frameworks for culpability.

More recently, Danaher (2016) refers to this issue as the retribution gap, describing it as a conflict between the principles of fairness and proportionality, which are central to retributive justice, and the inability to identify deserving targets of blame and punishment. He explains, "Fairness and proportionality are key aspects of the retributive philosophy: you give people what they deserve, nothing more or less. … When you combine a general desire to find appropriate targets of retributive blame, with the fact that no such targets can be found, you get a retribution gap" (p. 305). For Danaher, the retribution gap arises from a mismatch between the strong human inclination to blame and punish wrongdoers and the absence of suitable agents who justly warrant retributive blame.

[4] This notion of autonomy in the context of AI aligns with Hellstrom's (2013) gradient-based definition, which evaluates the autonomy of AI in terms of degree. According to this perspective, the level of an AI's autonomy is determined by the extent to which it can perform a diverse range of actions across various environments independently of human supervision. The greater its capacity to operate in this manner, the higher its degree of autonomy.

Similarly, Santoni de Sio and Mecacci (2021) underscore this concern, arguing that the culpability gap in the context of AI results from AI systems complicating prediction and control, thereby "creating *new legitimate reasons/excuses* for wrongdoing [my italics]" (p. 1061). This raises profound ethical challenges. Given the rapid development of AI and its increasing deployment across various domains, the practice of assigning responsibility remains critical for upholding justice for perpetrators, victims, and society. Moreover, responsibility attribution plays a vital role in regulating societal behavior by incentivizing adherence to moral and social norms. If, as many scholars claim, humans cannot be held responsible for actions resulting from AI's autonomy, significant ethical dilemmas are likely to emerge. How should we navigate and respond to this pressing issue? The next section will address this challenge in greater detail.

2.3 Why the Gap in Culpability is, at worst, less troubling than assumed or, at best, non-existent.

In this section, I synthesize some conceptual resources from the literature on this debate and integrate them with the principle of Meaningful Human Control (MHC) to argue that the culpability gap can be resolved and, ultimately, that it does not exist.

At the heart of the problem discussed in the previous section lies the concept of control[5]. Conceptually, whether the culpability gap exists or not depends on how the concept of control is understood. To clarify this point, I draw on the framework of "responsible AI through conceptual engineering"[6] proposed by Himmelreich and Köhler (2022). This framework distinguishes between two types of control: *strong* and *weak conception*. The former requires full and direct control, while the latter demands only a partial or indirect form.

This distinction raises an important following question: "What conception of control, if any, is required for the ascription of responsibility?" (Himmelreich & Köhler, 2022, p. 60). Thinkers such as Matthias appear to adopt a strong conception of control. This reliance on the strong conception justifies their argument for the existence of a culpability gap. However, I will contend that while the strong conception of control is sufficient for ascribing responsibility, it is not necessary. Instead, I will argue that a mere weak conception of control is *enough* to ground responsibility ascription[7]. Moreover, implementing this weaker conception aligns with, rather than being incompatible with, a *sense of justice* which Matthias refers to.

[5] I would like to set aside the knowledge condition as a means of addressing the culpability gap, primarily because human decisions to deploy AI in scenarios involving clear risks to life—such as autonomous vehicles (AVs) or autonomous weapons systems (AWS)—should, in principle, be foreseeable. This foreseeability implies that knowledge of potential risks can reasonably be assumed in such cases.

[6] Conceptual engineering is a methodological approach focused on the systematic evaluation and improvement of the concepts we employ, grounded in the normative assumption that these concepts can—and should—be refined to better serve collectively recognized normative goals. In this framework, the meaning and application of a concept are not fixed or immutable; rather, they are subject to deliberate reconsideration and revision, reflecting the idea that it is fundamentally "up to us" to determine how we construct and employ our conceptual frameworks.

[7] While some perspectives argue that "control" is not a necessary condition for responsibility—such as proponents of moral luck who contend that we sometimes bear responsibility for outcomes beyond our control—in this context, I assume that control is a necessary condition. If we create systems that we cannot control at all, there is little justification for building them in the first place. Thus, in principle, humans should maintain some form of control over AI systems.

I propose two ways to challenge the plausibility of this alleged gap. First, it is necessary to evaluate and investigate such cases individually by asking whether the situation that purportedly creates a responsibility gap actually warrants blame in the first place. In other words, is there truly a fault or a blameworthy action involved? The answer is obvious: not all cases of harm necessarily involve culpability, as there may be circumstances in which the relevant agent lacks intent, is not negligent, and faces an uncontrollable tragic accident. Such cases can be conceptualized as *blameless harm* (Hindriks & Veluwenkamp, 2022), which can be formulated by the following conditions:

1) The relevant agent lacks intention.
2) The agent is neither negligent nor reckless.
3) The harm results from a tragic accident entirely beyond the agent's control.

When these three conditions are met, the situation constitutes one of what we call blameless harm, where it is unwarranted to demand that "someone" should be blamed. A well-known example illustrates this point: imagine a driver traveling at a lawful speed and fully attentive. Suddenly, a pedestrian recklessly darts across the road, leaving the driver no time to react, resulting in a fatal accident. In this scenario, the driver neither intended to cause harm nor acted negligently or recklessly. Instead, the tragedy occurred due to the pedestrian's unpredictable actions, which were beyond the driver's control. While the driver may understandably feel self-reproach in this tragic situation, it would be neither fair nor reasonable to assign blame to him. This exemplifies a case of blameless harm, where no one is rightfully blameworthy, even though the outcome is deeply regrettable[8].

Nevertheless, proponents of the responsibility gap might recognize the possibility of blameless cases but still insist that in certain contexts—such as the use of AI in morally sensitive domains like warfare or medicine—blame should be assigned when the harm is morally unacceptable, given its impact on life or human dignity (Sparrow, 2007). But again, they argue that the autonomy of AI systems, operating beyond human control, creates situations in which responsibility cannot be justly or appropriately attributed, thereby sustaining the alleged gap.

This insistence leads to my second challenge to the plausibility of this alleged gap. Now, I propose a weak conception of control that does not require full and direct control in the way that proponents of the responsibility gap conceive it. To develop this approach, I draw upon the framework of Meaningful Human Control (MHC)[9]. This framework conceives the interaction between AI and humans as a socio-technical system—a decision-making unit comprising humans, AI, the physical environment, and the institutional infrastructure of society (Siebert et al., 2022). Within this perspective, human decision-making mechanisms and processes are not confined to the mind or brain; rather, they can extend to technologies and artifacts, such as AI (Santoni de Sio & van de Hoven, 2018).

In line with this, Nyholm (2017) introduces the concept of collaborative agency (or domain-specific supervised and deferential principled agency), which characterizes AI systems as entities capable of making specific decisions autonomously but requiring human collaboration. Such systems operate under human supervision, where humans initiate or reverse commands, oversee operations, and retain the authority to intervene or halt the AI's

[8] This case is comparable to the case of self-driving cars.
[9] I will elaborate on the MHC framework in the next section, but in this section I will only discuss some ideas relevant to the debate with proponents of the culpability gap.

actions at any point. Consequently, responsibility for AI actions remains with the human agents involved in the system. As Nyholm explains:

> The mere presence of unpredictability and a *lack of direct control* [my italics] are not by themselves enough to create responsibility-gaps. … Rather, … The humans involved are responsible for what the robots do for the reason that they initiate, and then supervise and manage, these human-machine collaborations. (p. 1217)

Combining both frameworks, we can say that humans and AI do not act *in isolation* but form integral parts of a unified decision-making system. Most importantly, human agents within this system still maintain and secure authority status and control through a *design stance that ensures responsiveness to human moral reasons*—a principle referred to by MHC theorists as the *tracking condition*.

In other words, the MHC approach conceptualizes "control" in a *top-level manner*, which does not require direct oversight of every detail of the system but allows for the delegation of specific functions to different parts, each performing its designated role (Di Nucci & Santoni de Sio, 2014). Simultaneously, relevant agents must design the system's overall behavior to align with human intentions, values, and moral reasoning. As Santoni de Sio explains:

> A human A may be in control of an action performed by an autonomous system S provided that S is part of a system Y whose general functioning is guided by the moral and practical reason of A; in particular, the distribution of tasks to different parts of the system is such as to allow the system to be responsive to the relevant moral and practical aims of A. (2016, p.13)

This perspective suggests that AI can remain under meaningful human control even when it operates autonomously, insofar as specific functions are delegated while the system's overall behavior is still responsive to human moral reasons. Therefore, when AI actions result in morally unacceptable harm, it reflects a failure on the part of relevant human agents—such as designers, software developers, manufacturers, operators, or others in enabling roles (Hindriks & Veluwenkamp, 2022, p. 21), supervisors (Nyholm, 2017), or commanders (Himmelreich, 2019).

By adopting this MHC-based understanding of control, we ensure, at least theoretically, that in cases of morally harmful outcomes—where someone is justifiably blameworthy—human agents can, in principle, be identified and held culpable. Understood as such, we can eliminate the gap by anchoring culpability within a structured socio-technical framework that aligns AI functions with human moral responsibilities. In this way, morally harmful outcomes can be attributed to human failings rather than to an unassignable "gap" in responsibility.

However, the practical implementation of the MHC approach in AI design and development is undeniably complex, requiring careful consideration of numerous factors that vary across social, political, cultural, institutional, legal, and application contexts. Each AI domain presents unique variables and challenges that must be addressed. For instance, questions arise such as: What moral values or principles should guide the design process? How should conflicting moral values be reconciled and integrated? How can we differentiate between blameless harm and the negligence of relevant agents? Additionally, how should

responsibility be distributed among the many actors involved in the development and deployment of AI systems?

While these are critical questions, addressing them in full lies beyond the scope of this article. The purpose of this section is to demonstrate, in principle, that an expanded understanding of control—grounded in the MHC approach—provides a robust framework to challenge the claim that a culpability gap exists. Importantly, this extended conception of control aligns with and does not contradict the moral framework for ascribing responsibility.

Nevertheless, this conceptual resolution is only part of the whole story. While it provides a theoretical solution to the culpability gap by emphasizing the authority, role, and capacity of humans as fully moral agents to design and govern AI systems in alignment with moral reasons, it abstracts away from the practical complexities. In reality, human agents involved do not always act with full rationality when interacting with AI. Automation bias, negligence or complacency, carelessness, and fatigue are discernible factors of human tendency that can diminish their willingness and abilities to take meaningful control and responsibility.

These issues pose a deeper challenge: not merely finding who is to blame but pointing to human failures to fully *recognize* and *embrace* their roles and responsibilities within socio-technical systems, leading to compromising safety standards and possibly resulting in life-threatening and socially hazardous events on a large scale.

This challenge extends beyond culpability and implicates two other forms of responsibility: accountability and active responsibility. These dimensions will be explored in the next section.

3. Accountability and Active Responsibility: Why they matter and what is the problem?

3.1 Accountability

Moral accountability, in its generally accepted sense, denotes the obligation of moral agents to provide explanations or justifications for their decisions and actions when questioned. This responsibility is backward-looking (or retrospective), typically arising in response to the "why question." Such questions may originate from various sources: a supervisor asking about an employee's tardiness, a professor querying a student's missed deadline, a victim of a traffic accident demanding an explanation for negligent driving, or the public scrutinizing potentially dubious policies enacted by politicians.

Furthermore, accountability holds instrumental value. As noted by Santoni de Sio and Mecacci (2021): "The process of exchanging questions and reasons helps finding explanations for things that have happened, reinforces trust and social connections between agents, and by exposing persons to potential requests for explanation and justification, it also tends to reduce undesired behavior by pushing persons to be more clearly aware of the impact of their actions on others and therefore motivated to prevent unwanted outcomes" (p. 1064)

Nonetheless, there is a deeper story as to why accountability holds significantly intrinsic rather than merely instrumental value. To explore this, I draw on the work of John Gardner, a legal philosopher, who illuminates this idea in his writings. His insights reveal how accountability is essential for moral agency and ownership of actions:

> As rational beings we cannot but aim at excellence in rationality. The only way we have to question that aim—by asking 'What reason do I have to excel at rationality?'—already concedes the aim by demanding a reason, by demanding that the case for rationality be made rationally. … as rational beings we cannot but want our lives to have made rational sense, to add up to a story not only of whats but of whys. (p.158)

Suppose we accept the idea presented above. It implies that the practice of accountability serves not merely instrumental values—such as avoiding culpability or fostering trust and social connections—but is also rooted in our very nature as rational beings. Each individual inherently yearns for reasons to make sense of their actions. In the act of taking accountability, individuals assert, develop, and cultivate their sense of moral agency and ownership of their actions. Gardner terms this "basic responsibility," from which the other form of responsibility, which he calls "consequential responsibility," including the obligation to explain one's actions and culpability (both legal and moral), is derived.

I would like to draw attention to the essential nature of accountability in its basic sense. The core of accountability here is not about whether others will ascribe or hold us accountable but rather about the individual *taking* and *exercising* accountability for themselves by using reason to reflect and understand their thoughts, decisions, and actions. In this process, two important elements are presupposed: first, we are moral rational agents, capable of reflecting on and making deliberate moral reasoning through our actions; second, there is a sense in which the outcomes of these actions *belong to* or are *owned by* the agent (van de Poel & Sand, 2018: 4473). Experientially, the agent perceives and identifies an action as their own by linking it to the reason underlying that action. As John Searle put it (2007), "We have the first-person conscious experience of acting on reason" (p. 57), which I will term *sense of ownership*.

In my view, the two presuppositions above form the foundation that makes the practice of ascribing and exercising responsibility in any form within society possible. More concretely, in the case of ascribing responsibility, I blame or hold you culpable and accountable for cheating on me because I presuppose that you have the morally rational capacity (which qualifies you as a moral agent) to resist temptation, even if it was intentional or due to a failure to act according to good reason.[10] Therefore, it is justifiable for me to blame you, and you, as a moral agent, can exercise responsibility by taking ownership of that action, which may include any form of expression such as accepting guilt, apologizing, expressing regret, or giving me some sort of excuse.

Before exploring how AI impacts and presents issues of accountability in the previously discussed sense, let me address another type of responsibility that is closely related and crucial for the ethics of AI and technology: active responsibility.

3.2 Active Responsibility

Active responsibility refers to a forward-looking type of responsibility. Unlike culpability and accountability, which deals with past actions, it concerns an agent's duty and

[10] This article does not aim to establish moral criteria for determining whether an action is right or wrong, which partially depends on socio-cultural contexts. Rather, the focus lies on addressing the implications for responsibility that follow *if* an action is deemed wrong. Moreover, my intention is merely to highlight that humans, as rational beings, possess the capacity to reflect on moral reasoning or consider "what ought to be," thereby qualifying as moral agents. I do not engage here with metaethical questions concerning what renders a particular normative or moral judgment true.

obligation to act according to societal and moral values, fulfilling obligations, roles, or duties proactively. This responsibility is connected to the agent's *engagement* with societal goals and multiple normative demands, focusing on their active participation in addressing moral and technical challenges and achieving objectives. It can be understood as "responsibility-as-obligations," where agents are expected to act based on norms and responsibilities assigned to them by society or their roles in an institutional realm (van de Poel & Sand, 2018).

It is especially crucial in the context of innovation and technology development, including AI. Designers, developers, engineers, and other involved agents bear significant responsibility to create technology ethically. Technologies are not value-neutral; they often embed and propagate certain values that can conflict with others. For instance, industrial reliance on fossil fuels may prioritize efficiency and productivity but harm sustainability, or smart energy meters might infringe on privacy rights, demonstrating the need for ethical foresight in technological advancements. These proactive ethical considerations are essential for ensuring that technology aligns with broader societal goals and moral values (Pesch, 2015; Unruh, 2000; Van den Hoven, 2013).

However, designing technology that reconciles and integrates the diverse values and needs of society, as well as those of stakeholders, is inherently complex. This complexity arises from the networked nature of engineering processes itself and the lack of clear "accountability forums" and rule systems akin to institutional structures found in government (Pesch, 2015, p. 6). As Santoni de Sio and Mecacci (2021, p. 1067) observe, engineers often lack a clear and consistent understanding of their social roles or shared systems of norms and rules to guide their professional decisions.

With the ambiguity surrounding the obligations designers should fulfill, another crucial characteristic of active responsibility becomes relevant in the process of designing and developing technology: responsibility-as-virtue (van de Poel & Sand, 2018; Williams, 2008). This form of responsibility does not concern specific actions or outcomes but emphasizes the character traits of agents. These include a willingness to take responsibility, an awareness of diverse normative demands within society, and the ability to exercise discretion and judgment appropriately when faced with conflicting value claims.

The significance of responsibility-as-virtue lies in its disposition to address gaps in role obligation that may arise even within a clearly defined division of roles and responsibilities in a collective project. Such gaps can occur due to unforeseen factors or changing circumstances, which may disrupt existing schemes of cooperation. Additionally, agents involved might fail to act in alignment with the responsibilities assigned to them. In these situations, what is needed is an agent with a virtuous sense of responsibility—one who is willing to take a proactive stance and demonstrate readiness to address unfulfilled responsibilities, even if those responsibilities do not strictly fall within their designated role. This virtue enables the agent to step forward and respond to the demands of the situation, ensuring that gaps in accountability are bridged.

Moreover, in cases where mistakes are made, an agent with responsibility-as-virtue will recognize and acknowledge that others have the moral right to hold them accountable for the consequences of their actions. This readiness to be responsible reflects their *morally agential capacity* to engage with legitimate moral reactions from others and take ownership of their actions.

Having established a comprehensive understanding of the importance of accountability and active responsibility, we can now turn our attention to the challenges posed by AI on human moral agency. This examination will focus on how these technologies influence humans' capacity for moral reasoning, which is a precondition for exercising accountability, and their ability to foster virtuous dispositions, essential for navigating socio-technical systems' ethical complexities.

3.3 The Accountability and Active Responsibility Problems: The Ironic Story of Moral Rational Beings in Interaction with AI

Earlier, I have highlighted the status of humans as moral agents endowed with the capacity for moral reasoning and the potential for virtuous dispositions. However, the reality is far less idealistic. Empirical research points to the profound psychological impacts that interactions with AI have on human (moral) agency. For instance, "automation bias" describes scenarios in which humans over-rely on or trust AI's decisions, often disregarding alternative sources of information, including their judgment (Soltanzadeh, 2024). Similarly, "automation complacency" refers to the attitude or belief that a system is sufficiently reliable to be left to operate and make decisions autonomously, even in high-stakes environments. This results in a deterioration of the human ability to detect errors—errors that could have ethical or moral consequences—due to a lack of vigilance and active monitoring (Zerilli et al., 2019). Another significant issue is "deskilling," where prolonged reliance on AI systems leads to a gradual decline in human abilities (Borgman, 1984). This impacts the role of humans along with their responsibility as operators, supervisors, and regulators, diminishing their capacity to handle risks and respond effectively to errors. Over time, this dependency erodes critical skills necessary for managing and mitigating failures within socio-technical systems.

Furthermore, an even more pressing and relevant issue is that AI not only diminishes technical skills and knowledge related to various professions (which may have ethical ramifications as well), but it also has the potential to erode our *moral skills*. Vallor (2014), for instance, examines the possible effects of new information and communication technologies (ICT) in fostering *moral deskilling*. She highlights cases where technologies such as robot caregivers may attenuate human attentiveness and care for others or where autonomous weapon systems may impair soldiers' sensitivity to morally salient features of specific battlefield situations. A similar concern is raised by Coeckelbergh (2013), who argues that military technologies like combat drones can weaken moral responsibility and empathy toward other human beings, ultimately making killing easier. He points out that such technologies facilitate what he terms "screenfighting," a mode of combat that not only increases physical distance but also creates moral-psychological distance between the killer and their target. As a result, the *embodiment* of the opponent fades from the perception of the one who kills, detaching them from the moral weight of their actions.

Drawing on virtue ethics as a moral framework, Vallor (2014) argues that the development and cultivation of moral skills, which are fundamental to virtuous character, require individuals to "engage repeatedly in the kinds of practices… that successfully engender certain skills of acting rightly in particular moral contexts" (p. 109). These practices must occur under the right conditions and provide opportunities for repetition**;** without these elements, moral dispositions cannot properly flourish**.** A crucial question that Vallor raises—one that is particularly helpful for evaluating AI and automation systems—is whether these technologies obstruct opportunities for moral skill development. If they do, then moral deskilling is likely to occur. I align with Vallor's evaluative framework and propose that accountability and active responsibility (as elaborated earlier) can themselves be

understood as moral skills. Like other virtues, these skills require appropriate conditions and repeated practice to be effectively cultivated. Building on this, I would like to further point out that prolonged interaction with AI as an intelligent automation system has the potential to deteriorate these moral skills over time.

To understand these phenomena, I draw on the highly relevant observations made by Cummings (2014), who synthesized these insights from extensive empirical research and highlighted their implications in his work. His remarks are as follows:

> Because of the inherent complexity of socio-technical systems, decision support systems are particularly vulnerable to certain potential ethical pitfalls that encompass automation and accountability issues. If computer systems diminish a user's sense of moral agency and responsibility, an erosion of accountability could result. In addition, these problems are exacerbated when an interface is *perceived as a legitimate authority.* (p. 23)
>
> and
>
> Automated decision support tools are designed to improve decision effectiveness and reduce human error, but they can cause operators to relinquish a sense of responsibility and subsequently accountability because of a perception that the automation is *in charge*. [My italics] (p.25)

Ironically, as Banks et al. (2018, p. 283) aptly point out, AI is "most dangerous when it behaves consistently and reliably for most of the time." The reason is that as AI becomes increasingly advanced and technically sophisticated, humans begin to perceive it (both knowingly and unknowingly) as an *independent agent capable of making decisions and exercising responsibility on its own*. This perception leads humans to *experientially distance* themselves from the outcomes of the system's actions, thereby failing to exercise accountability or take ownership of the AI's actions. When humans abdicate, as it were, this role and dismiss these skills due to misplaced trust in the system's reliability, it can result in undesirable or even catastrophic outcomes.

The challenges discussed thus far might seem overly pessimistic, but they are grounded in a substantial body of empirical evidence accumulated over years of research. These challenges are also likely to intensify with the continued advancement of AI. While it is impossible to address all these issues comprehensively here, one crucial ethical concern that deserves emphasis is *the negative impact of AI on the realization of our rational capacities in operation and, by extension, our self-understandings and awareness of ourselves as morally responsible agents within socio-technical systems.*

Ideally, humans, as genuine, responsible agents, should retain the capability to oversee, monitor, and intervene in AI operations when necessary. However, the reality often seems to be the reverse: AI exerts significant influence over human perception, attitudes, and decision-making processes due to the factors outlined earlier. This phenomenon has given rise to what I term the problem of *disengagement*. AI's growing reliability and perceived authority increasingly distance humans from actively engaging with their moral responsibilities. As Coeckelbergh (2016, p. 753) puts it:

> Conditions for responsibility seem to require that we engage with our environment and others, with the human and non-human environment. Ethics, it seems, is about such engagement. *A failure to act responsibly is a failure to engage*.

More precisely, human agents who disengage from AI's operational automation experience a deterioration in the conditions required for exercising or taking responsibility (though not necessarily for holding it). As previously discussed, this erosion affects two fundamental presuppositions for basic accountability: their status as rational beings capable of articulating reason for action and their sense of ownership over actions and decisions.

While societies or institutions may establish and enforce rules to attribute responsibility in terms of culpability, by virtue of formal laws and principles, accountability and active responsibility are more than that; they require agents to feel responsible *from within*. Without this internal sense of responsibility, there emerges what can be identified as an *accountability gap* and *active responsibility gap*, even if the agent can still justifiably be blamed (e.g., for carelessness or negligence) in terms of culpability. Thus, the absence of genuine moral engagement and the lack of perceived ownership on the part of the human agent are critical barriers to achieving meaningful accountability and active responsibility.

One must concede that addressing this issue poses significant challenges. It demands more than mere conceptual refinement, such as that undertaken in discussing culpability (Section 2.3). Rather, it calls for the interdisciplinary integration of knowledge from fields such as engineering, design, sociology, humanities, and human factors—a branch of psychology focused on creating systems that accommodate human limitations and optimize human capabilities. While the specifics of such interdisciplinary approaches fall beyond the scope of this article, leveraging the framework of Meaningful Human Control (MHC), which I have touched upon earlier, offers a potential way to *work around* (not solve) these challenges. In the final section, I will attempt to propose approaches to mitigating some aspects of the problem by elaborating on the "tracing conditions" for MHC.

3.4 Re-construction of the Problem and Preliminary Guidance

Before elaborating on the Meaningful Human Control (MHC) framework, I would like to refine the framing of the problem and offer some insights that can clarify its nature and provide an initial direction for addressing it.

As said above, human interaction and reliance on AI not only influence perceptions, thoughts, attitudes, and decisions about tasks in which humans are involved but significantly affect *our self-understanding and the relationship between ourselves and AI within the socio-technical system*. This implies that how we perceive or believe about AI fundamentally determines how we interact with it and, more specifically, shapes our self-perception regarding our role in our interactions with AI (as mere passively disengaged monitors or as actively engaged collaborators).

The starting point for mitigating the challenges posed by AI is not to naively assert human superiority based on the notion that technology is a value-neutral "tool" serving rational, autonomous human beings. Instead, we must acknowledge that AI possesses some degree of agency (Soltanzadeh, 2024) and has the potential to influence society and culture at large (Segessenmann et al., 2023). Particularly, it is crucial to recognize AI's foundational role in shaping human lifestyles, actions, experiences, and decisions—or even non-decisions—in a world increasingly dependent on AI across various domains. However, this acknowledgment

is not about succumbing to AI's influence but about understanding the risks it poses to fundamental human values, particularly responsibility, which define our humanity. One of the most pressing risks is how AI's influence can subtly undermine human moral agency and accountability.

To clarify, my proposal to acknowledge AI's role in shaping our self-understanding as moral agents—and the associated risk to the value of responsibility—does not necessarily lead to a dystopian scenario of losing autonomy, which shatters moral practice and is often perceived as an irrecoverable outcome. On the contrary, recognizing the impact of AI on us opens a valuable opportunity to reassess the roles of "we as humans" and "AI as machines," along with the relationships between them within the socio-technical system. Such recognition encourages us to reestablish conceptual foundations in terms of ontology and anthropology, delineating clear boundaries between what precisely humans and AI are and what kind of relationship they (should) have. This involves defining the appropriate roles and statuses of both entities within the system while identifying factors that may challenge these boundaries. The goal is to navigate this dynamic landscape effectively to avoid losing our sense of responsibility or ability to act responsibly while also ensuring the appropriate use and development of AI. Ultimately, it is humans who create meaning and define narratives. As Mark Coeckelbergh (2021) aptly describes through the concept of "narrative or hermeneutic responsibility":

> AI is part of our narratives and helps to shape them, but it is our responsibility to define their role as co-creators and it is up to us what place and role we give them in the narratives that we co-create. And in the end, it is our responsibility to decide what narratives we want to (co-)write—including narratives about AI, with AI, and sometimes against AI. (p.2446)

I fully agree with the above perspective and wish to emphasize and empower humans as meaning-makers of the world around them. One conceptual framework that I find particularly suitable for revisiting and defining the relationships and roles between humans and AI is the approach of Meaningful Human Control (MHC). This framework will be elaborated upon in the next section.

4. Meaningful Human Control

The Meaningful Human Control (MHC) framework, as a systematic philosophical account proposed by Santoni de Sio and van den Hoven (2018), integrates three key concepts: "Guidance Control," which grounds moral responsibility in the capacity for rational control over actions (Fischer & Ravizza, 1998); and "Responsible Innovation," emphasizing the ethical development of technology; and "Value-sensitive Design," focusing on embedding societal and moral values into the creation of socio-technical systems. These systems, comprising humans, AI, the physical environments in which they operate, as well as legal frameworks and institutions, are designed to align with social and moral values. This alignment is achieved through inclusive deliberations among all stakeholders, aimed at negotiating and refining a set of socially and ethically desirable values. These values are then translated into "normative design requirements," which are explicitly embedded in the design and development processes from the outset. By incorporating such principles early in the development cycle, this approach seeks to preemptively address potential ethical challenges, mitigating adverse impacts while ensuring that AI systems are designed to serve societal needs and values properly.

The core purpose of MHC is to ensure that humans continue to function as morally responsible agents within socio-technical systems. This framework guarantees that AI does not become "out of control" in ways that might create responsibility gaps (Section 2.2). This is achieved by expanding the concept of control, not through a requirement for direct control, but rather by designing the system's overall behavior to align with human values, reasons, and intentions. This requirement is referred to as the "tracking condition"[11] (Section 2.3). In addition, MHC emphasizes fostering active responsibility among involved agents. This entails encouraging them to proactively engage with and improve their understanding of the system's capabilities and limitations. It also involves motivating agents to remain aware of their roles, obligations, and responsibilities, particularly regarding the potential societal impacts of AI behavior. This aspect is encapsulated in the "tracing condition" (see below).

In the next section, I will elaborate on the tracing condition and demonstrate how it provides actionable guidance for mitigating issues of disengagement, which are deeply intertwined with questions of accountability and active responsibility (as discussed in Section 3.3).

Elaborating on the Tracing Condition

In the literature advocating for the Meaningful Human Control (MHC) framework, there is a consistent emphasis on the tracing condition as the design of a socio-technical system that enables the behavior of AI to be traced back to human agents involved in its design, development, and use. More importantly, there must be at least one human agent who:

> possesses both (i) sufficient knowledge of the system's capabilities and limitations of the system and (ii) sufficient moral awareness and the capacity to fulfill their role as a legitimate target of accountability for the system's behavior. … Tracing requires the alignment of the system with the capacities of the relevant human agents. (Santoni de Sio and Mecacci, 2021, p.1007)

Regarding the first requirement, we need human agents with both the *technical capacities* and *motivation* to understand the potential impacts of the system, actively monitor and supervise it, and intervene when necessary, appropriately and safely. For the second requirement, we need *moral capacities* and *motivation* characterized by responsiveness to relevant moral reasons, as well as the ability and willingness to fulfill one's duties within one's role. This includes going beyond their designated role as well when ethical implications demand it, thereby taking active responsibility to address potential issues. The importance of designing systems that enable traceability between the behavior of AI and the relevant human agents lies not only in facilitating the identification of those who should be held culpable (or blamed) in cases of unwanted outcomes. It also plays a critical role in *re-engaging* human agents experientially with their roles and moral responsibilities in interacting with AI, empowering them to effectively exercise those responsibilities.

To foster the technical capacities and motivation of humans involved in the use and development of AI, it is insufficient to focus solely on technical understanding of the system. We must also consider the *reasonableness* of task allocation between AI and humans. This requires insights from psychology regarding motivation, as well as the cognitive capacities

[11] For those interested in a detailed elaboration and practical operationalization of the tracking condition, including a case study on designing self-driving cars, see the article "Meaningful Human Control as Reason-Responsiveness: The Case of Dual-Mode Vehicles" (Mecacci & Santoni de Sio, 2019).

and limitations of both humans and AI, to design "interfaces" that align with these characteristics. These ideas are actively explored in fields such as human factors, human-machine systems, and engineering (Zerilli et al., 2019; Cavalcante Siebert et al., 2022), which aim to develop socio-technical systems that balance the strengths and limitations of humans and AI fittingly.

However, a greater challenge lies in fostering moral capacities and motivation in practice. This directly pertains to the moral skill of human agents to exercise accountability and active responsibility for development and moral cultivation. As previously discussed, addressing the issue of disengagement, which creates accountability and active responsibility gaps, requires more than merely designing legal and moral frameworks or normative principles to impose responsibility on relevant human agents. Instead, *it necessitates cultivating an internal sense of responsibility and awareness of the moral connection between one's own agency and the outcomes of AI's actions.*

The question is: *how can we cultivate a sense of responsibility within humans in the context of AI usage and development?* One approach I propose, though it may sound cliché, remains vital—developing educational programs aimed at training individuals to adopt an ethically responsible mindset. These programs should focus on fostering the moral character of human agents in the context of AI technology, emphasizing the cultivation of virtues such as care, responsibility, and moral excellence.

To advance an educational curriculum that integrates ethical aspects into AI development, relevant institutions—including organizations, governments, and universities—must design an environment that fosters ethical deliberation for both users and developers throughout their training. One effective approach would be to incorporate ethicists and philosophers into specific courses or specialized modules while also creating platforms for discussion on the complex societal impacts of AI (Grosz et al., 2019). Moreover, the curriculum should emphasize experiential learning, providing individuals with opportunities to develop ethical decision-making and action-oriented skills in realistic scenarios. An empirically qualitative study by Griffin et al. (2024), which surveyed 40 tech developers, found that while most expressed a strong interest in learning more about ethics, they were disengaged by traditional ethics courses, which are often structured as standalone lectures and fail to inspire motivation. Instead, they advocated for an integrated approach, where ethics education is embedded within hands-on learning experiences, guided by expert mentorship and real-world application.

However, in practice, designing an ethics education system that yields tangible and lasting impacts remains highly challenging**.** Various approaches have been proposed and piloted in AI ethics curricula within engineering and computer science programs. For instance, Northeastern University has incorporated an activity called "Values Analysis in Design" into AI-related courses (Kopec et al., 2023). Evaluation results indicated a positive shift in students' moral attitudes, particularly regarding ethical responsibility. However, the overall body of empirical research on the precise effectiveness of different pedagogical methods remains scarce, highlighting the need for further investigation (Hardebolle et al., 2025, p. 134).

One viable approach to fostering developers to recognize and exercise ethical responsibility is the "RESOVEDD strategy." This framework outlines nine structured steps that enable developers to engage in ethical decision-making in a reasoned and systematic

manner. Additionally, it allows them to justify and articulate the decision-making processes that inform their product design.

Drawing on an empirical study by Vakkuri et al. (2019), their team investigated whether incorporating this ethical tool into technology design processes could enhance ethical awareness. The study involved five groups of computer science students assigned to develop an innovation while integrating the RESOVEDD strategy as a design requirement. The findings indicated that this tool effectively fostered ethical considerations in practice, particularly in terms of transparency and a sense of responsibility. Although the study suggested that this approach alone may not yet cultivate enduring intrinsic motivation, I argue that it represents a promising starting point—one that can serve as an essential component in the ongoing development of developers' ethical competencies.

A strong sense of ethical motivation and commitment can be cultivated by helping developers recognize the significance and meaningful impact of their work on others and society. As Bowie (1998) has emphatically stated, experiencing a sense of meaningfulness in one's work can actively contribute to an individual's moral development. To facilitate such an experience, it is essential to provide resources that draw from narratives and case studies of exemplary individuals with virtuous character. These role models can serve as guides, helping individuals develop and refine their ethical sensibilities and cultivate a responsible moral character through deliberate practice and reflection.

Thus, the curriculum cannot solely focus on technical, engineering, or experimental psychology knowledge. It must draw on conceptual, narrative, and theoretical resources from the humanities—social and political philosophy, history, and literature—to encourage *practical wisdom* (phronesis) regarding core humanly moral values. Such an approach would inspire individuals to critically reflect on their roles and responsibilities when engaging with AI systems. The core of these programs would integrate humanities and engineering alongside other relevant disciplines, ensuring a holistic perspective on AI development and usage. Graduates from these programs would emerge as professionals equipped not only with technical expertise but also with the ethical and philosophical acumen to contribute meaningfully to processes of AI design, development, and public policymaking.

Beyond establishing educational programs for AI developers as professionals, AI ethics education for users and citizens is equally important. This is especially crucial for children and young people, who represent society's future and will grow into adulthood alongside the rapid advancements of AI. AI ethics should be integrated into K-12 curricula to lay a foundation for ethically aware citizens. This could be done by embedding AI ethics topics into various subjects (Touretzky et al., 2019) or by developing dedicated AI ethics courses (Burton et al., 2017). However, specific considerations regarding curriculum design, pedagogical approaches, and assessment strategies fall beyond the scope of this article and warrant independent research in their own right.

In addition to establishing formal and dedicated educational programs, building a responsible AI society requires integrating other crucial resources. These include driving initiatives within civil society, cultural domains, and the media, all of which play pivotal roles in upholding and promoting the value of responsibility. Such collective efforts could help create an "established accountability forum" within the public sphere, encouraging open dialogue, transparency, and reflection on the ethical implications of AI technologies. By integrating these diverse dimensions—educational, social, and cultural—the vision of a

responsible AI society becomes more attainable, grounded in a shared commitment to accountability and ethical progress.

Before concluding, I would like to leave a final reflection. While the approaches proposed here may face significant challenges in practice due to numerous constraints—such as economic disparities in less affluent regions, political systems characterized by centralized power that marginalize the voices of the public, or the profit-driven incentives of large corporations operating under capitalist market systems that often neglect ethical considerations—these factors collectively obscure and hinder the motivation and moral capacities of individuals to flourish. Moreover, there is a persistent tendency of humans to evade responsibility, particularly in cases where direct causal outcomes are attributed to AI. These obstacles undeniably present barriers to achieving meaningful ethical progress. However, this article aims to make such hidden and entrenched problems more visible, bringing them to the forefront of critical inquiry. I hope that this effort will inspire new perspectives, deeper understanding, and a renewed sense of hope, determination, and commitment to collectively shaping *our* future in the age of artificial intelligence.

Conclusion

As humans increasingly interact with and rely on AI, their sense of responsibility diminishes. This occurs because humans often perceive AI as an independent, reliable agent capable of exercising responsibility on its own, leading to a failure to exercise responsibility and take ownership of outcomes. I argued that establishing external rules, frameworks, or principles to impose responsibility on relevant human agents is insufficient. Accountability and active responsibility require an internal sense of responsibility, which cannot be externally mandated.

To work around this problem, I drew upon the framework of Meaningful Human Control (MHC), emphasizing the design of socio-technical systems that align with core human values. I highlighted the necessity of educational programs that cultivate virtue, responsibility, and practical wisdom in the context of AI use and development. Beyond the education system, however, societal change requires the active participation of civil society, media, and cultural practices, all of which should uphold and celebrate the value of responsibility. This multifaceted effort is essential to preserving the status of humans as morally responsible agents, ensuring that AI does not supplant us in defining and upholding our own agency.

REFERENCES

Banks, V. A., Plant, K. L., & Stanton, N. A. (2018). Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016. *Safety Science, 108*, 278–285. https://doi.org/10.1016/j.ssci.2017.12.023

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Bowie, N. E. (1998). A Kantian Theory of Meaningful Work. *Journal of Business Ethics, 17*, 1083–1092. https://doi.org/10.1023/A:1006023500585

Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). Ethical considerations in artificial intelligence courses. *AI Magazine, 38*(2), 22–34. https://doi.org/10.1609/aimag.v38i2.2731

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., Van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: actionable properties for AI system development. *AI Ethics, 3*, 241–255. https://doi.org/10.1007/s43681-022-00167-3

Coeckelbergh, M. (2016). Responsibility and the moral phenomenology of using self-driving cars. *Applied Artificial Intelligence, 30*, 748–757.

Coeckelbergh, M. (2023). Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making. *AI & Society, 38*(6), 2437–2450. https://doi.org/10.1007/s00146-021-01375-x

Cummings, M. (2014). Automation and accountability in decision support system interface design. *The Journal of Technology Studies, 32*(1). https://doi.org/10.21061/jots.v32i1.a.4

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics Inf Technol, 18*, 299–309. https://doi.org/10.1007/s10676-016-9403-3

Di Nucci, E., & Santoni de Sio, F. (2014). Who's afraid of robots? Fear of automation and the ideal of direct control. In Battaglia, F., & Weidenfeld, N. (Eds.), Roboethics in Film. Pisa University Press.

Doorn, N. (2012). Responsibility ascriptions in technology development and engineering: Three perspectives. *Science and Engineering Ethics, 18*(1), 69–90. https://doi.org/10.1007/s11948-009-9189-3

Floridi, L., & Sanders, J. (2004). On the Morality of Artificial Agents. *Minds and Machines 14*, 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Fischer, J., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Gardner, J. (2003). The mark of responsibility. *Oxford Journal of Legal Studies, 23*(2), 157–171. https://doi.org/10.1093/ojls/23.2.157

Goetze, T. (2022). Vicarious responsibility in AI systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22) (pp. 390-400). ACM. https://doi.org/10.1145/3531146.3533106

Griffin, T. A., Green, B. P., & Welie, J. V. M. (2024). The ethical wisdom of AI developers. *AI and Ethics.* https://doi.org/10.1007/s43681-024-00458-x

Grosz, B. J., Grant, D. G., Vredenburgh, K., Behrends, J., Hu, L., Simmons, A., & Waldo, J. (2019). Embedded EthiCS: Integrating ethics across CS education. *Communications of the ACM, 62*(8), 54–61. https://doi.org/10.1145/3330794

Gunkel, D. J. (2017). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology, 22*(4), 307–320.

Hardebolle, C., Héder, M., & Ramachandran, V. (2025). Engineering ethics education and artificial intelligence. In S. Chance, T. Børsen, D. A. Martin, R. Tormey, T. T. Lennerfors, & G. Bombaerts (Eds.), *The Routledge International Handbook of Engineering Ethics Education* (1st ed., pp. 125–141). Routledge. https://doi.org/10.4324/9781003464259

Hanson, F. A. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and Information Technology, 11*(2), 91–99. https://doi.org/10.1007/s10676-009-9184-z

Himmelreich, J. (2019). Responsibility for killer robots. *Ethical Theory and Moral Practice, 22*(4), 731–747. https://doi.org/10.1007/s10677-019-10007-9

Himmelreich, J., & Köhler, S. (2022). Responsible AI Through Conceptual Engineering. *Philos. Technol, 35*, 60. https://doi.org/10.1007/s13347-022-00542-2

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: from responsibility gaps to control gaps. *Synthese*, *201*, 21. https://doi.org/10.1007/s11229-022-04001-5

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Köhler, S., Roughley, N., & Sauer, H. (2017). Technologically blurred accountability? Technology, responsibility gaps and the robustness of our everyday conceptual scheme. In *Moral agency and the politics of responsibility* (pp. 51–68). Routledge

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology, 24*(36). https://doi.org/10.1007/s10676-022-09643-0

Kopec, M., Magnani, M., Ricks, V., Torosyan, R., Basl, J., Miklaucic, N., Muzny, F., Sandler, R., Wilson, C., Wisniewski-Jensen, A., Lundgren, C., Baylon, R., Mills, K., & Wells, M. (2023). The effectiveness of embedded values analysis modules in computer science education: An empirical study. *Big Data & Society, 10*(1), 20539517231176230. https://doi.org/10.1177/20539517231176230

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1.

Mecacci, G., & Santoni de Sio, F. (2020). Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics Inf Technol, 22*, 103–115. https://doi.org/10.1007/s10676-019-09519-w

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility-loci. *Science and Engineering Ethics*, *24*, 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Pesch, U. (2015). Engineers and Active Responsibility. *Sci Eng Ethics, 21*, 925–939. https://doi.org/10.1007/s11948-014-9571-7

Santoni De Sio, F. (2016). Ethics and Self-driving Cars: A White Paper on Responsible Innovation in automated Driving Systems.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI,* 5, Article 15. https://doi.org/10.3389/frobt.2018.00015

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philos. Technol, 34*, 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Searle, J. R. (2007). *Freedom & neurobiology: Reflections on free will, language, and political power*. Columbia University Press.

Segessenmann, J., Stadelmann, T., Davison, A., & Dürr, O. (2025) Assessing deep learning: a work program for the humanities in the age of artificial intelligence. *AI Ethics, 5*, 1–32. https://doi.org/10.1007/s43681-023-00408-z

Soltanzadeh, S. (2025). A metaphysical account of agency for technology governance. *AI & Soc*, 40, 1723-1734. https://doi.org/10.1007/s00146-024-01941-z

Strawson, P. F. (1974). *Freedom and resentment*. In P. F. Strawson (Ed.), Freedom and resentment and other essays (pp. 1–25). Methuen.

Siebert, L. C., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI and Ethics, 3*(2), 241–255. https://doi.org/10.1007/s43681-022-00167-3

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019). Envisioning AI for K-12: What should every child know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 9795–9799. https://doi.org/10.1609/aaai.v33i01.33019795

Tigard, D.W. (2021). There Is No Techno-Responsibility Gap. *Philos. Technol, 34*, 589–607. https://doi.org/10.1007/s13347-020-00414-7

Unruh, G. C. (2000). Understanding carbon lock-in. *Energy Policy, 28*(12), 817-830. https://doi.org/10.1016/S0301-4215(00)00070-7

Vakkuri, V., & Kemell, K.-K. (2019). Implementing AI ethics in practice: An empirical evaluation of the RESOLVEDD strategy. Springer International Publishing. https://doi.org/10.1007/978-3-030-35151-7_21

Van de Poel, I., & Sand, M. (2021). Varieties of responsibility: two problems of responsible innovation. *Synthese*, *198*(Suppl 19), 4769–4787. https://doi.org/10.1007/s11229-018-01951-7

Van den Hoven, J. (2013). Value-sensitive design and responsible innovation. In R. Owen, J. Bessant, & M. Heintz (Eds.), *Responsible innovation* (pp. 75–83). Wiley. https://doi.org/10.1002/9781118551424.ch4

Van de Poel, I., Nihlén Fahlquist, J., Doorn, N., Zwart, S., & Royakkers, L. (2012). The problem of many hands: Climate change as an example. *Science and Engineering Ethics, 18*(1), 49–67. https://doi.org/10.1007/s11948-011-9276-0

Williams, G. (2008). Responsibility as a virtue. *Ethical Theory and Moral Practice, 11*(4), 455–470. https://doi.org/10.1007/s10677-008-9109-7

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic decision-making and the control problem. *Minds and Machines, 29*(4), 555–578. https://doi.org/10.1007/s11023-019-09513-7