Personal Autonomy and Autonomous Systems: Sameness and Difference

6

Fiorella Battaglia

Abstract

This chapter aims to clarify the transformative effects on the concept of personal autonomy resulting from extending the notion to Artificial Intelligence (AI) technologies, particularly in the defense domain. The complexity of this task is increased by the fact that philosophers have offered a wide range of competing accounts of the autonomous agent's special relation to their own desires and values. I will draw on the concept of autonomy implied in John Martin Fisher and Mark Ravizza's control theory, which, while explaining the theory and practice of autonomy, also identifies two conditions that cannot be appropriately satisfied by autonomous systems, but only by persons who interact and communicate with one another. Despite the superficial sameness between personal autonomy and autonomous systems, control theory allows us to pinpoint an important distance in their underlying features. The seeming similarity masks indeed a profound conceptual divergence between autonomy as a condition of moral agency and autonomy understood as the characterization of a system's operational capacities. This supports, at least in principle, the moral argument that entrusting machines with authority over human life and death is ethically wrong, regardless of how advanced the technology might be.

Keywords

Autonomy · Meaningful human control · Kant · Reactive attitudes and reasons · Epistemic injustice

F. Battaglia (✉)
Dipartimento di Studi Umanistici, University of Salento, Lecce, Italy

Ludwig Maximilians University of Munich, Munich, Germany
e-mail: fiorella.battaglia@unisalento.it

M. Anzalone et al. (eds.), *Reconfiguring Human Autonomy*, Integrated Science 40,
https://doi.org/10.1007/978-3-032-14537-6_6

1 Introduction

The concept of autonomy is key to understanding our relationship with Artificial Intelligence (AI). Autonomy is not merely one of many ethical concerns in the field of AI—rather, "it is the central philosophical problem when we confront the human-machine mutual connection". How we understand and apply the concept of autonomy in relation to machines shapes the entire ethical landscape surrounding AI. In this chapter, I argue that a precise conceptual analysis is needed to avoid misunderstanding of the term. By clarifying what autonomy truly entails—and what it does not—I aim to challenge assumptions about so-called "autonomous" systems and to provide a foundation for a more critical and philosophically informed debate on both the autonomy of machines and that of humans. A key requirement of autonomy is the control condition [1–3]. The control condition modifies our understanding of autonomy by emphasizing the normative dimensions of agency and responsibility. Therefore, the idea of control affects how we understand autonomy and—as a consequence—it is crucial for understanding our relationship with AI. Contrary to views that treat control as a secondary or derivative feature of agency, the concept was introduced precisely because autonomy alone may be insufficient for grounding responsibility. Without a coherent account of control, claims about moral or legal responsibility risk collapsing into uncertainty [4]. This notion loses its technicality and takes on a profile that differentiates human and machine control. It helps to disambiguate the concept of autonomy and therefore to make its profiles clear. Therefore, it can be said that control clarifies the morally relevant dimensions of autonomy: accountability, answerability, and responsiveness to reasons. The concept of control has undergone significant refinement and has emerged as the more nuanced concept of *meaningful human control* (MHC), capturing the essential role of human agency in overseeing autonomous systems [5, 6]. When articulated in terms of MHC, it strikes at the core of what renders AI ethically controversial.

2 Meaningful Human Control (MHC)

Importantly, the concept of meaningful human control (MHC) has not only shaped the debate on lethal autonomous weapon systems (LAWS) but has also proven indispensable in critically assessing the ethical challenges posed by other autonomous technologies, including self-driving cars and surgical robots. This expansion highlights MHC's essential role in defining the moral boundaries of delegating control to machines in various sensitive areas of human action. Concerns about control offer a broad framework for understanding the ethical challenges associated with AI, while also highlighting the unique issues that these technologies introduce. An autonomous car crash causing a passenger's death is perceived by the media and public not as a routine incident, but as an ethically significant event. Somehow, since AI is involved, the incident is supposed to have greater moral significance. While the outcomes are the same, people tend to perceive moral significance differently if the controlling agent is a machine rather than a human.

I believe that the beautiful yet unsettling figure of Maria—an image that captures the essence of the uncanny valley phenomenon [7], where a machine appears eerily "almost but not quite human" and evokes profound discomfort—can help explain the unease we experience when confronted with a lethal machine. There is a pre-reflective, intuitive grasp of the distinction between us and other entities that either appear human-like or act in ways that mimic human behavior. This immediate response plays a central role in shaping how we ascribe moral status and interpret the ethical significance of relational practices. In the case of humanoid machines or highly sophisticated AI, this recognition often manifests as a form of discomfort—a reaction aligned with what has been described as the uncanny valley. What unsettles us is not simply that these entities resemble or act as humans, but that they disrupt the structure of mutual recognition on which our ethical practices unfold. As such, our moral unease does not stem merely from novelty or unfamiliarity, but from a more profound disruption of the second-personal structures—of address, response, and accountability—that underlie ethical life [8–10].

This is exactly the reason why the debate on control [2, 11, 12] has shifted from concerns about control in general to a critical focus on MHC. Before the rise of autonomous systems, discussions naturally focused on forms of human control, making it unnecessary to specify the nature of the controlling agent. However, with the advent of a mixed population of agents—both human and artificial—the need has arisen to explicitly distinguish the type of agent exerting control. The concept of MHC emerged as a critical response to the ethical challenges posed by LAWS. The debate about MHC in the context of LAWS has centered on the crucial question of whether such systems should be granted control over life-and-death decisions.

The notion that a machine could autonomously select and lethally harm human targets without human oversight has faced strong objections from a broad range of scholars [13–15]. However, it is important to note that some scholars counter this position by arguing that it would be unethical to forgo the use of such technologies, particularly when they have the potential to reduce human casualties and increase precision in combat [16, 17]. This debate highlights the complex ethical issues surrounding lethal autonomous weapon systems, [18] and underscores the need for a careful consideration of control, responsibility, and humanitarian outcomes. Amoroso and Tamburrini in their 2020 paper argue for a threefold role for human control on weapon systems to be "meaningful" and explain that this would allow humans to have a more active and influential role in the use of weapons [14]. First, humans must occupy the critical role of 'failsafe actor,' tasked with intervening to prevent or mitigate harmful malfunctions in autonomous weapon systems. This responsibility underscores the indispensable need for human oversight to safeguard against unpredictable errors and ethical breaches, reinforcing the argument that MHC is essential for maintaining accountability and moral legitimacy in the deployment of such technologies. Second, human control must be structured in a way that ensures that responsibility can be clearly attributed, thereby preventing accountability gaps—situations in which it becomes unclear who is responsible for wrongful actions [19]. Such gaps arise when autonomous systems, which lack moral agency and legal personhood, make decisions independently, making it difficult or

impossible to hold any human or institution accountable [20]. Addressing these gaps is crucial to maintaining the integrity of legal and ethical frameworks and ensuring that wrongful acts do not go unaccountable. Third, human control must guarantee that decisions affecting the life, physical integrity, and property of individuals remain firmly in human hands rather than being delegated to machines. This is seen as ethically crucial as such decisions carry profound moral weight and require judgment, empathy, and accountability, which are not to be found in machines [21]. Moreover, international legal frameworks, including humanitarian and human rights law, emphasize the need for human responsibility to ensure that the use of force complies with principles of proportionality, distinction, and necessity. Delegating these decisions to autonomous systems risks eroding these safeguards, creating ethical blind spots and legal ambiguities that could lead to unjust outcomes and violations of fundamental rights.

3 From a More Technical Standpoint: Navigating Deontological and Consequentialist Ethics

As previously mentioned, Arkin adopts a contrasting stance, arguing that it would be unethical to forgo the use of autonomous systems [16, 17]. His position is supported primarily from a technical standpoint. According to him, the primary reasons for employing autonomous, robotic, or unmanned systems on the battlefield include the following. First, autonomous and robotic systems enable a smaller number of soldiers to accomplish missions that previously required significantly larger forces. By enhancing individual soldier capabilities, these technologies effectively increase military effectiveness and operational efficiency. Second, robotic systems enable combat operations to be conducted over significantly larger areas than was previously feasible. Third, robotics empower individual soldiers to extend their reach deeper into the battlespace, for instance, by enabling them to observe or engage targets at greater distances. Finally, robotic systems allow for the reduction or removal of soldiers from direct combat roles.

After having presented the theoretical reasons, Arkin moves on to offer four compelling justifications for the deployment of autonomous systems—emphasizing the alleged benefits they promise. First, autonomous systems can potentially reduce human casualties by taking soldiers out of direct harm. Second, they may increase the precision of targeting, thereby minimizing collateral damage and civilian deaths. Third, such systems can operate tirelessly and with consistent effectiveness, unlike humans who may suffer from fatigue or emotional stress. Finally, autonomous systems could enhance compliance with the laws of armed conflict by making more calculated and unbiased decisions based on pre-programmed ethical guidelines. Although these points possess a certain degree of argumentative force, LAWS have faced many criticisms and objections in both public and academic discourse. In this regard, there are three primary ethical objections [19]. First and from a more empirical standpoint, based on our current understanding of technological development, robots in the foreseeable future will not possess the capacity to make the complex

practical and moral distinctions required by the laws of armed conflict—such as distinguishing between combatants and non-combatants, assessing proportionality in the use of force, and determining the military necessity of violent action. As a result, delegating military tasks to autonomous systems risks increasing the occurrence of wrongful acts and violations of international humanitarian law during military operations. Second, on more principled grounds, it is morally wrong to allow a machine to exercise control over decisions concerning the life and death of human beings, regardless of how advanced the technology may be. According to this position, such applications are considered *mala in se*: intrinsically wrong by their very nature, irrespective of context or outcome. This stance rests on the belief that certain moral decisions, particularly those involving the taking of human life, must remain within the realm of human agency, where accountability, empathy, and moral deliberation are possible. Delegating such decisions to autonomous systems, which lack consciousness, intentionality, and moral responsibility, constitutes a fundamental violation of human dignity. The two positions reflect a deeper ethical divide, with one rooted in deontological principles—emphasizing duties and moral absolutes—and the other grounded in consequentialist reasoning, focusing on outcomes and overall harm reduction. From a deontological perspective, allowing machines to make life-and-death decisions undermines the moral status of persons by treating them as objects of calculation rather than subjects of moral concern. Even if such systems could act more precisely or efficiently, the act itself remains ethically impermissible. This view stands in contrast to consequentialist arguments—such as those advanced by Ronald Arkin—which focus on outcomes, suggesting that if autonomous systems can reduce harm or increase compliance with the laws of war, their use may be not only permissible but also morally required [17, 22].

The debate, then, hinges on whether moral permissibility is determined solely by outcomes or whether certain roles, especially those involving lethal force, are inherently reserved for moral agents. From a more theoretical standpoint, the involvement of autonomous weapon systems in military operations may overcomplicate, or even render impossible, the attribution of moral and legal responsibility in cases of war crimes or fatal accidents [19]. In recent years, the ethical concerns surrounding autonomous weapon systems have been distilled into a guiding principle within the legal-political debate, as reflected in Article 36 [23]. The Principle of MHC asserts that future weapon systems must ensure that humans, rather than computers or algorithms, retain ultimate control over the use of (lethal) force [15]. This means that human agents should remain morally responsible for all significant decisions related to (lethal) military operations. Given the central conceptual outcome of the chapter, the following sections aim to make this insight more intelligible through practical elaboration and contextual application. In particular, to render the principle of MHC operational, it is necessary to articulate its requirements from both technical and legal perspectives. Only by grounding the principle in concrete mechanisms of system design and established legal norms can it effectively guide the development and governance of autonomous technologies.

4 Moral Theory, Control and Autonomy

Santoni de Sio and van den Hoven [5] provide an analysis of the concept of "control" grounded in the philosophical literature on free will and moral responsibility, where control is a central theme. Their perspective is shaped by the longstanding debate in moral theory concerning compatibilism and incompatibilism. Compatibilism and incompatibilism no longer play a central role in moral theory as they once did. Nevertheless, it is important to acknowledge that the paper by Santoni de Sio and van den Hoven contributes to the so-called compatibilist theory of moral responsibility, which holds that individuals can be held accountable even if their actions are causally determined, potentially within a deterministic framework. More importantly, the paper demonstrates how MHC can be systematically designed for, thereby extending the value-sensitive design approach within the ethics of technology to a new and critical domain. There are alternative approaches that, while strengthening certain aspects of MHC, remain insufficient to fully meet the comprehensive set of requirements necessary for MHC to be adequately ensured. Being physically or operationally present in the control loop does not, in itself, ensure that a human is meaningfully guiding or responsible for the system's actions. This mere proximity to the system is insufficient, as a person may be present and able to influence certain aspects of the system through causal intervention, while remaining unable to affect other components—some of which may be more ethically or operationally significant. Even exercising control in the form of making a substantive causal contribution may be insufficient to qualify as MHC. As a matter of fact, the conditions for MHC closely resemble the most stringent standards for legal responsibility. Santoni de Sio and van den Hoven argue that agents must not only possess causal control over events but also satisfy stricter conditions related to knowledge, intention, capacity, and opportunity.

The account of control—and, connected with it, of autonomy—can be situated on a metaethical level [24]. Incompatibilists hold that causal explanations are fundamentally incompatible with genuine autonomy and moral responsibility. This view is reflected both in theories that posit a metaphysical realm of freedom to safeguard the notion of free will and in those that reject free will altogether by extending causal determinism to all events. Contemporary libertarians (e.g., [25–27]), following a philosophical tradition dating back at least to Immanuel Kant, maintain that certain human beings possess a distinctive form of autonomy often referred to as "contra-causal" freedom. This capacity grounds moral responsibility and autonomy in a unique metaphysical status, establishing them in a way that no other kind of being can plausibly claim. From a metaethical perspective, compatibilists believe that humans can be morally responsible for some of their actions, even if they lack any special metaphysical power to escape the influences that cause their behavior. Recent developments in compatibilist thought [1, 2, 7, 28], adopt a more nuanced view of the human mind and agency. They reject the notion that mere mental causation is sufficient to ground autonomy and moral responsibility. Instead, they emphasize the importance of an agent's capacity for rational control over their actions—specifically, the ability to reflect upon, evaluate, and act in accordance

with reasons—as the foundation of moral responsibility. There is some kind of sympathetic alignment with current accounts, such as Strawson's [8] and Darwall's [9, 10], highlighting the ways in which our moral practices are deeply rooted in interpersonal relationships and social expectations. People often perform actions that others may not consider morally justified [8, 29]. When challenged, individuals typically respond by offering reasons they believe justify their actions. In doing so, they engage in a process of assuming and negotiating responsibility. Importantly, this process is not isolated or purely internal—it is socially situated. In sum, responsibility is embedded in (1) the roles individuals occupy within social institutions, (2) the continuity of their past actions and character, and (3) the cultural norms and moral expectations that define what counts as acceptable behavior. Similarly, autonomy, intrinsically bound to the notion of responsibility, resists reduction to the detached frameworks typical of purely technical interpretations.

Ultimately, these considerations highlight how Arkin's account, although technically robust, fails to address the more nuanced aspects of the relationship between autonomy and responsibility that make the ethical design of autonomous systems inseparable from the concept of meaningful human control. Consequently, the proposal to treat the deployment of unmanned systems on the battlefield as a kind of categorical imperative is met with an equally substantive and coherent counterargument.

5 The Second-Person Standpoint

This relational view of responsibility aligns with what Darwall [9] has called the second-person standpoint: an ethical framework that sees moral obligations as arising not from detached judgment, but from interpersonal demands made between moral agents. Responsibility, in this view, presupposes the capacity for mutual recognition, communication, and answerability—capacities that machines do not possess. This has direct implications for debates on artificial agency and accountability. Unlike humans, autonomous systems cannot participate in second-person moral practices: they cannot recognize others as moral equals, respond to blame with justification or remorse, or internalize shared norms. As a result, attributing moral responsibility to machines risks bypassing the deeply relational structure of human accountability. If responsibility presupposes this kind of mutual moral engagement, then current AI systems—no matter how advanced—lack the requisite autonomy for bearing moral responsibility in the same sense as human agents. Robert Brandom's inferentialist philosophy, particularly developed in *Making It Explicit* [30], argues that the ability to participate in practices of giving and asking for reasons is foundational to what it means to be an autonomous, responsible agent. For Brandom, meaning is not simply a matter of representing the world, but of being embedded in social practices where claims are made, challenged, justified, and assessed. This discursive interaction defines our status as sapient beings—creatures who can not only follow norms but also justify their actions and beliefs by appealing to shared standards. Brandom directly links this practice to responsibility, arguing

that to assert something or to act within a normative space is to undertake a commitment and accept accountability to justify it. The key aspect in his argument is that the capacity to enter justificatory relations is not merely cognitive or computational—it is social and normative. One must be able to recognize others as reason-givers and reason-demanders, and in turn be recognized by them.

This has profound implications when applied to the ethics of artificial intelligence and autonomy. AI systems, no matter how advanced, do not participate in these normative discursive practices. They do not take responsibility for their assertions or actions as Brandom understands that, because they do not recognize themselves—or others—as participants in a community of practice. They process inputs and produce outputs, but they do not enter commitments, respond to challenges, or modify their positions based on mutual understanding or moral reflection. From this standpoint, it becomes clear that autonomy in the human, normative sense cannot be attributed to machines simply because they operate independently. Without the capacity for taking and giving reasons, they lack the core of what Brandom identifies as autonomous agency—and by extension, they cannot bear moral responsibility in any robust sense [31]. Brandom's emphasis on the normative practice of giving and asking for reasons underscores a fundamental limitation of autonomous systems in ethical decision-making. Clearly, machines cannot enter the mutual relationships of recognition and accountability that confer moral significance on decisions. It follows that MHC requires humans to retain ultimate authority and responsibility over decisions involving life, death, and moral judgment precisely because humans are the only agents capable of participating in these norm-governed, discursive practices. MHC thus reflects the necessity of preserving a second-person standpoint—a relational context where agents hold one another accountable through reasons, justifications, and mutual recognition. Machines, lacking this capacity to engage in normative discourse, cannot assume the role of responsible agents. They cannot justify their actions or respond to moral demands; instead, they function as causal devices devoid of moral accountability.

This framework challenges that idea of machine ethics which aims to design AI systems that can make ethical decisions autonomously. Without the capacity for genuine reason-giving and taking, machine ethics remains a form of simulated morality, a set of programmed rules and responses rather than authentic moral agency. Consequently, the ethical control of AI systems must remain in human hands—not only to ensure responsibility but also because moral autonomy is intrinsically tied to the human capacity for normative engagement. In other words, Brandom's framework justifies the insistence on MHC as a conceptual and moral necessity: machines can assist or execute certain functions, but ultimate control and moral responsibility must rest with human beings who can participate in the dialogue of reasons that underpins ethical life.

Building on this, I intend to introduce another account that supports a second-person standpoint. While developed within the basic conceptual structure of the relationship between philosophy and neuroscience, the account by Ravizza and

Fisher is nevertheless capable of effectively engaging in the machine ethics debate. Their framework emphasizes the relational and normative dimensions of autonomy, making it particularly relevant for distinguishing between human autonomy and the autonomy exhibited by artificial systems. According to Fischer and Ravizza [2], in order for a person to be morally responsible for an action X, they must possess what the authors call "guidance control" over that action. This concept implicitly conveys a robust notion of autonomy, as it requires not only the capacity to respond to reasons but also a reflective identification with the mechanism by which decisions are made. The concept of guidance control, as developed by Fischer and Ravizza, provides a substantive account of what it means to act autonomously in a morally responsible way. Autonomy, in this context, is not simply a matter of acting independently or without external coercion; rather, it involves acting through a decision-making mechanism that is both moderately reason-responsive and owned by the agent. The first condition—moderate reason-responsiveness—links autonomy to the agent's capacity to recognize, evaluate, and act upon moral and practical reasons. This ensures that the agent's behavior is not merely the product of blind causation or internal compulsion but reflects responsiveness to normative considerations and relational practice. The second condition—that the mechanism be "the agent's own"—introduces a deeper, relational dimension to autonomy. For Fischer and Ravizza, an agent must not only possess a functional capacity for responsiveness but must also have taken ownership of the process through which decisions are made. This involves understanding how one's actions are generated and accepting responsibility for them, which in turn requires a degree of self-reflection and integration into a broader moral community.

Taken together, these conditions suggest that autonomy, properly understood, is not just a matter of control or choice, but of normative engagement and identification with one's reasons for acting. This conception contrasts sharply with simplified notions of autonomy often attributed to machines, where autonomy is equated with independent functioning rather than with moral responsibility or reflective ownership. Darwall [9] argues that moral obligation, accountability, respect for persons, and the distinctive freedom of moral agents are all irreducibly second-personal or relational. This means that these concepts cannot be fully understood from a detached, third-person perspective [32–34] or even a purely first-person point of view [35]. Instead, they arise within a space of reciprocal moral address, where individuals make claims, demands, and justifications to one another as equals in the moral community. From this standpoint, to say someone is morally obligated is not merely to describe their behavior or attitudes, but also to state that others have standing to hold them accountable, to demand reasons for their actions, and to expect certain forms of acknowledgment or redress. Responsibility, then, is not a solitary function of inner conscience or rational deliberation alone—it depends on the interpersonal structure of recognition and answerability.

6 Relational Autonomy

Some significant implications for the concept of autonomy can be pinpointed. In Darwall's framework, autonomy is not limited to self-governance or internal coherence; it also involves the capacity to engage in mutual moral relations, respond to others' demands, and acknowledge the authority of shared moral norms. Autonomy, in this sense, is inherently relational and dialogical. Applied to contemporary debates in AI ethics, this perspective underscores that autonomous systems, no matter how sophisticated, are fundamentally incapable of genuine second-personal engagement. They do not recognize others as moral equals, nor can they issue or respond to moral demands in a way that reflects genuine accountability or respect. As such, they cannot be full participants in moral practices grounded in Darwall's second-person standpoint—reinforcing the argument that moral responsibility and autonomy remain uniquely human capacities. Brandom, Strawson, and Darwall collectively support a conception of autonomy that is not simply about individual self-governance or independence, but instead it emphasizes the importance of our involvement in a normative community where agents hold each other accountable. According to this perspective, autonomy is exercised through participation in practices of giving and asking for reasons [30], being subject to and issuing reactive attitudes like blame or resentment [8] and recognizing oneself and others as bearers of moral authority and obligations [9]. Relational theorists of autonomy argue that appropriate social relationships are something more than just supportive of autonomy: they are fundamental to its very structure.

Following this line of thought, autonomy should not be conceived merely as internal self-governance, but rather as a capacity that is cultivated, acknowledged, and sustained within a network of social and interpersonal relationships. Scholars such as Oshana [36], Mackenzie [37], and Nedelsky [38] emphasize that autonomy requires conditions of mutual recognition, respect, and access to genuinely meaningful choices. Without these, even agents who meet internal criteria may lack the standing or opportunity to act autonomously in any robust sense [39]. In this light, being appropriately related to others—being heard, taken seriously, and afforded fair options—is not a causal condition that helps us to interact with others, but rather a pivotal element for autonomous agency. Such a relational perspective challenges the idea of the autonomous individual as isolated and self-contained, and instead places autonomy within the very fabric of social life. This insight into the relational foundations of autonomy resonates with accounts of language and normativity that similarly recognize the social character of agency. In this regard, Brandom defends a conception of discursive practice in which language is structured normatively through what he calls its "downtown": the game of giving and asking for reasons. At the heart of this practice is the idea that to speak is to engage in a normative activity—advancing claims, drawing inferences, and navigating a network of commitments and entitlements. When speakers make assertions, they undertake commitments that can be challenged or defended through reasons. If successfully justified, these commitments earn entitlement. Importantly, this is not just one language game among many; rather, it is constitutive of discursive practice itself.

Without this normative structure, there would be no meaningful linguistic communication at all. Rather than being a solitary or inward capacity, autonomy emerges through second-personal engagement which unfolds as recognizing others as sources of legitimate moral claims and being recognized in turn.

The idea of relational autonomy, articulated in this way, effectively demonstrates the radical difference between human autonomy and machine autonomy—a difference that this chapter has sought to highlight. It also shows that the moral significance of our actions is inseparable from the shared practices, expectations, and mutual recognition that constitute our social world. Autonomy gains its depth and normative force precisely because it is embedded in the meaningful webs of interpersonal connection—and machines cannot impose themselves as leading agents within this uniquely human sphere. Recognizing this fundamental distinction has important implications not only for the design of autonomous systems but also for how we conceive of the human-machine relationship itself. As technological capabilities continue to advance, it remains essential to remember that authentic moral agency depends on forms of engagement and accountability no machine can replicate. Although artificial systems may appear increasingly like us in their operations, there are radical differences that must be clearly acknowledged, to avoid human autonomy being conflated with artificial autonomy—and ultimately risk being overshadowed by it.

7 Conclusion

The conception I have been arguing for situates autonomy within a relational and dialogical space, where moral agency depends on responsiveness to norms that are shared, upheld, and negotiated between persons. It does not reject Kant's emphasis on autonomy as a necessary presupposition of moral obligation and the dignity of persons. However, it offers a distinct interpretation of this idea by emphasizing that these moral concepts contain an irreducibly second-personal dimension. In this view, autonomy is not simply the capacity for self-legislation of a rational will but is grounded in our ability to relate to others as mutually accountable moral agents—agents who have reactive attitudes and can issue, recognize, and respond to moral claims within a shared normative space. This sharply contrasts with the algorithmic or operational autonomy of artificial systems, which lacks the relational dimension of human autonomy. Machines may act independently, but they do not possess the capacity to recognize or respond to the moral authority of others. As such, they cannot be autonomous in the second-personal sense, nor can they bear the moral responsibilities that follow from such autonomy.

Competing Interests This work was partially supported by the project "Ethical Design for AI", part of the Spoke 6 of the project FAIR–Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and by the European Union–Next Generation EU, Missione 4 Componente 1 CUP F53D23010740001 DigitHuman.

References

1. Dennett DC (1984) Elbow room: the varieties of free will worth wanting. MIT Press, London
2. Fischer JM, Ravizza M (1998) Responsibility and control: a theory of moral responsibility. Cambridge University Press, New York
3. Shepherd J (2015) Deciding as intentional action: control over decisions. Australas J Philos 93(2):335–351
4. 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. RoboLaw series. Pisa, Pisa University Press, pp 127–144
5. Santoni de Sio F, van den Hoven J (2018) Meaningful human control over autonomous systems: a philosophical account. Front Rob AI 5:1–14
6. Robbins S (2024) The many meanings of meaningful human control. AI Ethics 4:1377–1388. https://doi.org/10.1007/s43681-023-00320-6
7. Dennett DC (2014) When AI kills who is to blame? In: Battaglia F, Mukerji N, Nida-Rümelin J (eds) Rethinking responsibility in science and technology. Pisa University Press, Pisa, pp 203–214
8. Strawson P (1963) Freedom and resentment. Proc Br Acad 48:187–211
9. Darwall SL (2006) The second-person standpoint. Harvard University Press, Cambridge
10. Darwall SL, Dill B (2014) Moral psychology as accountability. In: D'Arms J, Jacobson D (eds) Moral psychology and human agency: philosophical essays on the science of ethics. Oxford University Press, New York, pp 40–83
11. Mele AR (1995) Autonomous agents: from self-control to autonomy. Oxford University Press, New York
12. Greenspan P (1978) Behavior control and freedom of action. Philos Rev 87:225–240
13. Asaro P (2012) On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. Int Rev Red Cross 94(886):687–709. https://doi.org/10.1017/S1816383112000768
14. Amoroso D, Tamburrini G (2020) Autonomous weapons systems and meaningful human control: ethical and legal issues. Curr Robot Rep 1:187–194
15. Véliz C (2021) Moral zombies: why algorithms are not moral agents. AI & Soc. https://doi.org/10.1007/s00146-021-01189-x
16. Arkin R (2008) Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture. In: Proceedings of the 3rd ACM/IEEE international conference on human–robot interaction. Association for Computing Machinery, New York, pp 121–128. https://doi.org/10.1145/1349822.1349839
17. Arkin R, Ulam P, Wagner AR (2012) Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. Proc IEEE 100:571–589
18. Blanchard A, Chris T, Taddeo M (2023) Ethical governance of artificial intelligence for defence: normative tradeoffs for principle to practice guidance. Available at SSRN: https://ssrn.com/abstract=4517701 orhttps://doi.org/10.2139/ssrn.4517701
19. 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
20. List C, Pettit P (2011) Group agency: the possibility, design, and status of corporate agents. Oxford University Press, Oxford/New York
21. Yıldız T (2025) The minds we make: a philosophical inquiry into theory of mind and artificial intelligence. Integr Psychol Behav Sci 59:10. https://doi.org/10.1007/s12124-024-09876-2
22. Wagner A, Borenstein J, Arkin R (2019) Kantian one day, consequentialist the next: moral emotions as mediators between ethical frameworks for robots. In: Proceedings of the International Association for Computing and Philosophy (IACAP), June 2019

23. Art 36 (2015) Killing by machine: key issues for understanding meaningful human control. http://www.article36.org/autonomous-weapons/killing-by-machine-key-issues-for-understanding-meaningful-human-control/
24. Abbink D, Amoroso D, Cavalcante Siebert L, van den Hoven J, Mecacci G, Santoni de Sio F (2024) Introduction to meaningful human control of artificially intelligent systems. In: Research handbook on meaningful human control of artificial intelligence systems. Edward Elgar Publishing, Cheltenham, pp 1–11
25. van Inwagen P (1983) An essay on free will. Oxford University Press, New York
26. Kane R (1996) The significance of free will. Oxford University Press, New York
27. Hodgson D (2012) Rationality + consciousness = free will. Oxford University Press, New York
28. Frankfurt HG (1971) Freedom of the will and the concept of a person. J Philos 68(1):5–20. https://doi.org/10.2307/2024717
29. Korsgaard CM (1992) Creating the kingdom of ends: reciprocity and responsibility in personal relations. Philos Perspect 6:305–332. https://doi.org/10.2307/2214250
30. Brandom R (1994) Making it explicit: reasoning, representing, and discursive commitment. Harvard University Press, Cambridge, MA
31. Baum K, Mantel S, Schmidt E et al (2022) From responsibility to reason-giving explainable artificial intelligence. Philos Technol 35:12. https://doi.org/10.1007/s13347-022-00510-w
32. Snow CP (1959) The two cultures and the scientific revolution. Rede lecture. Cambridge University Press, Cambridge
33. Pereboom D (2001) Living without free will. Cambridge University Press, New York
34. Greene J, Cohen JD (2004) For the law, neuroscience changes nothing and everything. Philos Trans R Soc B 359(1451):1775–1785
35. Nagel T (1986) The view from nowhere. Oxford University Press, New York
36. Oshana M (2006) Personal autonomy in society. Ashgate, Aldershot
37. Mackenzie C (2000) Imagining oneself otherwise!: toward a feminist theory of relational autonomy. In: Mackenzie C, Stoljar N (eds) Relational autonomy: feminist perspectives on autonomy, agency, and the social self. Oxford University Press, New York. https://doi.org/10.1093/oso/9780195123333.003.0007
38. Nedelsky J (2011) Law's relations: a relational theory of self, autonomy, and law. Oxford University Press, Oxford
39. Fricker M (2007) Epistemic injustice: power and the ethics of knowing, Online edn. Oxford Academic. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001. Accessed 6 July 2025

Fiorella Battaglia is the Founding Director of the Laboratory for Ethics in the Wild at the Digital Humanities Centre, University of Salento, where she is also associate professor of moral philosophy in the Department of Humanities. Her research focuses on challenging ethical questions resulting from emerging technologies and climate change, which shape both our social and epistemic practices and our moral experiences. After obtaining her MA degree in Philosophy from the University of Pisa, she earned her PhD in Philosophy and Politics from the University of Naples "L'Orientale" (2004). In 2016, she completed her habilitation in Practical Philosophy and received her venia legendi from the Ludwig-Maximilians-Universität in Munich (Germany). She has also held an assistant professorship of Social Philosophy at the Berlin-Brandenburg Academy of Sciences and Humanities, at the Humboldt University in Berlin, an adjunct professorship of Epistemology at the Faculty of Medicine of the University of Pisa, and a visiting professorship at the Dirpolis and Biorobotics Institutes of the Sant'Anna School of Advanced Studies in Pisa.