What does it take to be a moral agent?

Emma Brambilla[1]

Abstract. The rise of increasingly sophisticated AI systems displaying what appear to be agentive behaviours has renewed the debate on whether artificial agents should or should not qualify as moral agents. Much of this discussion has focused on identifying the properties that are traditionally associated with moral agency, including consciousness, autonomy, free will and control over one's actions. This paper argues for a shift in perspective: while these features may play an important role in our understanding of human agency, none of them seems fit to provide stable ground for determining whether AI systems qualify as moral agents. Rather than asking whether AI systems possess the "key ingredients" traditionally associated with moral agents, we should focus on the normative practice of attributing responsibility. Drawing on Wilfrid Sellars' (1956) notion of the Space of Reasons, I argue that moral agency should not be understood as primarily dependent on as a set of internal capacities, but rather as participation in a "game" of mutual accountability. Even if contemporary AI systems may come close to satisfying some of the criteria associated with functional accounts of agency, functional agency alone would still not establish moral agency. To develop this claim, I introduce the notion of "symmetrical vulnerability", arguing that the practices of reciprocal accountability presuppose that participants are mutually exposed to demands for justification, blame and normative sanction. The ethical challenge posed by AI, therefore, is not whether machines should be recognized as moral agents, but rather about the complex redistribution of human responsibility within increasingly complex scenarios. The task of AI ethics is not to extend moral agency to artificial systems, but to reconceptualize human responsibility under conditions of distributed control.

1 INTRODUCTION

The lightning fast development of increasingly sophisticated AI systems has radically intensified the philosophical debates around the nature of moral agency. LLMs and other AI systems seem now capable of performing tasks that, until a few decades ago, appeared to require distinctly human capacities. As these systems become embedded into everyday social, medical, economic and political settings, questions concerning their ethical status have become increasingly urgent. Much of the debate approaches the issue by focusing on the properties (or "key ingredients") traditionally associated with morally responsible agents. Discussions often centre around whether AI systems showcase, or could eventually display, consciousness, genuine autonomy, free will or control over their actions. The underlying assumption of this line of inquiry is that determining the presence or absence of such faculties in AI systems would ultimately settle the question at hand. Yet, this strategy continues to face a number of difficulties. Not only are many of the capacities listed above highly controversial even in human cognition, but there is also little to no agreement on which of them should be considered really necessary for moral responsibility. The result is a debate that frequently stagnates between competing conceptions of agency which, given the pace at which AI systems are getting implemented, we cannot afford. This paper argues that the problem may have been tackled in the wrong way. Rather than asking whether AI systems possess a particular set of internal properties, I suggest we shift our attention on the normative practice of responsibility attribution. Following this intuition, moral agency should no longer be understood exclusively as a matter of possessing certain capacities. Instead, it should be considered a crucial part of the social and normative practices within which agents hold one another accountable. Building on the notion of the Space of Reasons[2], I argue that moral responsibility presupposes participation in relations of mutual accountability. To develop this claim, I endorse the notion of Symmetrical Vulnerability[3]. This is not to be intended as mere physical vulnerability, but as a practice of holding one another responsible that primarily requires participants to be mutually exposed to demands for justification, criticism, blame and normative sanctioning. A moral agent is one that can occupy both ends of the normative relationship of accountability. The paper is articulated as follows: Section 2 will examine the limits of what I shall call the "ingredients approach" to moral agency. Section 3 will clear the distinction between functional agency and moral agency. Section 4 introduces the normative framework of responsibility developed through the concept of the Space of Reasons. Section 5 develops the notion of symmetrical vulnerability, arguing it constitutes a necessary condition for participation in practices of moral accountability. Finally, Section 6 considers the implications of this account for responsibility gaps.

[1] Post graduate student, University of Florence.

[2] Sellars, 1956.

[3] Vallor & Vierkant, 2024.

2 THE LIMITS OF THE "INGREDIENTS APPROACH"

Discussions of human moral agency have traditionally proceed by identifying a set of faculties thought to be constitutive of moral agency. When we think of ourselves as moral agents, we tend to attribute to us a set of capacities: this is what I will be referring to as the "ingredients approach to moral agency". Within AI ethics, the term "ingredients approach" does not indicate that contemporary theories simply reduce moral agency to a checklist of properties, but it is aimed to capture a broader tendency to approach the question of artificial moral agency by asking whether AI systems instantiate the relevant features that appear to ground moral responsibility. To create a moral agent, the recipe lists as follows: as sprinkle of consciousness, complete autonomy, a decent amount of control and overall freedom. Imagine one sits down and actively tries to gather all the ingredients: the first instinct would be to check if we already have everything we need in our pantry. Then, the question becomes the one about whether AI systems also possess them or not. It is common in machine ethics to distinguish between types of moral agents on the basis of how advanced their moral capacities are: the very aim of AI is to model or simulate cognitive capacities[4]. Among the most eligible ingredients for human agency we can find consciousness[5], autonomy and free will[6], intentionality and control[7]. The point I want to stress before continuing is that these capacities are far from irrelevant. However, their philosophical status, as I shall briefly illustrate, remains deeply contested even in the case of human agents. This implies that their role in grounding moral responsibility in regards to AI systems may be less straightforward than it may seem.

Starting with control, at least in regards to human agency, moral responsibility is often linked with the ability of an agent to control the consequences of his actions[8]. However, control can be understood in a number of different ways, already making it hard to regard it as an essential ingredient. A first understanding of the word could imply the agent has the ability to choose among different paths. However, as shall be briefly explored in the next session, this understanding mixes control with multiple-choice freedom which, given the open debate around compatibilism and incompatibilism, is far from settled. Another understanding of control is prospective, in the sense that the agent can predict the consequence of their actions and act accordingly. Yet, and this is pretty uncontroversial, complete control on the consequences of a determined action is arguably never available to human agents. Absurdly, if full control was a requirement for moral responsibility, ordinary human agents would frequently fail to qualify as morally responsible[9]. Against this, the fact that moral responsibility keeps being meaningful proves that control cannot serve as a necessary ingredient for AI agency either.

The search continues, but soon finds a similar difficulty with both freedom and autonomy. The one around free will is possibly one of the most ancient philosophical debates, and yet it is not even close to reaching its conclusion. The debate stagnates around two main groups: the compatibilists, who think freedom and determinism are somehow reconcilable, and incompatibilists, who disagree. A recent survey[10] has shown that the majority of philosophers are more sympathetic towards compatibilisit solutions. Most contemporary compatibilist accounts, such as the one developed by Daniel Dennett, understand agency primarily in terms of reason responsiveness, rational deliberation and goal-directed behaviour. Functionalist approaches to human agency such as Dennett's have the advantage of avoiding some of the metaphysical difficulties associated with free will. Yet they tend to fail short when leaving a door open for a blurring of the distinction between functional agency and moral agency[11]. The shrinking of intentionality to a mere predictive strategy seems to open the door for the possibility that AI systems could potentially satisfy the autonomy or deliberative criteria (though this point is highly questionable). However, even if this was true, the question about moral responsibility would remain open. Even though Dennett's understanding of freedom is far from being the only one[12], it captures well the problem of mixing functional agency with moral agency, a topic that I shall briefly address in the next section.

The appeal to consciousness appears, at first sight, more promising. Many philosophers regard phenomenal consciousness as an internal ability of living beings[13] and at the same time as a necessary condition for moral agency. Therefore, the apparent absence of conscious experience in AI systems, given that, contrary to humans, they do not have biological bodies, could be taken to settle the debate. However, this ingredient is also destined to prove problematic. For one, there is little to no agreement

[4] Misselhorn, 2022.
[5] Chalmers, 1996; Searle, 1980; Nagel 1974.
[6] Dennett, 1984; Fischer & Ravizza, 1998: Porter, 2024.
[7] Floridi & Sanders, 2004.
[8] Fischer & Ravizza, 1998; Frankfurt,1971.
[9] Dennett, 1984; Dennett, 1987; Dennett, 2003.
[10] https://survey2020.philpeople.org/survey/results/4838
[11] Dennett & Caruso, 2021.
[12] Pereboom, 2022.
[13] Seth, 2025.

concerning the nature of human consciousness itself, with some accounts still confusing consciousness with intelligence. Furthermore, the debate is even more sparse when questioning under which conditions it might emerge in AI systems. As philosophers, we should be trained to avoid setting foot on slippery grounds, at it is precisely for this reason that the question about consciousness, while it need not cease existing, is a much too unstable criterion to resolve the pressing matter at hand. I want to make clear that the intent of this paper is not in any way aimed at diminishing the importance of the debate around AI consciousness, which is of crucial importance. What I hope has since now emerged is not that consciousness, autonomy, free will and control should be regarded as irrelevant for agency overall. Rather, my argument is perhaps more pragmatic, and it is aimed to show that none of these ingredients is, at the actual stage, stable enough to resolve the question about the possibility of morally responsible AI agents. This suggests the need for a shift in perspective, one that, I suggest, focuses not on internal properties, but on a normative practice that gets carried out on a daily basis.

Finally, I would like to spend a few words on why I think the ingredients approach is usually the first that comes to mind when tackling the problem of AI moral agency. The focus on the conversational abilities of large language models (LLMs), along with the very reference to "intelligent" systems, has contributed to a distributed tendency to project human features on AI systems, leading us to ultimately feel not only inferior to them, but also to attribute them a moral status and intentional abilities from a surface-level performance[14]. This is precisely what Simone Natale[15] refers to when he talks about "deceptive machines", and what Luciano Floridi[16] defines "agency without intelligence". While most philosophers who specialize in the field of AI Consciousness and Ethics rightly deny that current AI systems display genuine agentive behaviors, behavioral indistinguishability is often taken by who stands outside of the debate as ontological equivalence. While linguistic sophistication does not necessarily entail consciousness, nor does simulated responsiveness amount to genuine intentional participation in the space of reasons (as we shall see in Section 4 & 5), I argue that the tendency to anthropomorphize technology is what also motivates at least part of the ingredients approach. It seems that the tendency to project human features of AI systems functions in two ways: from the outside-in, when we attribute them human capacities, and from the inside-out, when we try to establish if they effectively possess them.

3 FUNCTIONAL AGENCY DOES NOT EQUAL MORAL AGENCY

The inadequacy of the Ingredients Approach does not imply that AI systems lack agency altogether, and I shall now clarify why. Some contemporary AI systems seem to display behaviours that could be described as agentive: they pursue certain goals, adapt to changing environments, respond to inputs and have some degree of autonomy in determining action outputs. Even if most are very cautious in attributing strong agency to AI systems, several have argued that AI systems could at least satisfy functional accounts of agency. As briefly mentioned in the previous section, according to Dennett's functionalist account of freedom, agency does not necessarily require consciousness or metaphysical freedom. Rather, what matters is the ability of a complex system to exhibit a coherent and goal-directed behaviour. This is then explainable through the notion of intentional stance, a predictive strategy that treats other complex systems as if they possessed beliefs and desires. As Catrin Misselhorn notes: "Dennett makes it too easy for machines to be moral agents"[17].

Whether contemporary AI systems genuinely satisfy such accounts remains an open debate. However, for the purposes of this paper, the issue will be set aside. In fact, even if it was granted that current AI systems display a kind of functional agency, from this it would not follow that they qualify as moral agents. Functional and moral agency should not get mixed up as the distinction between the two has a prominent role in contemporary discussions. The risk about merging functional agency and moral agency is to no longer have an account of moral responsibility that is not merely a consequentialist one[18].

As mentioned in the previous section, the conversational sophistication showcased by the most recent LLMs often encourages overzealous attributions of intentionality, understanding and even responsibility to these systems. Empirical evidence shows that technologies can indeed create powerful illusions of agency[19], and contemporary AI systems have been described as exhibiting forms of "agency without intelligence"[20]. It is crucial that the apparent display of agency is not taken as an opening for a possible participation in the normative practices that underpin moral responsibility. Addressing this matter

[14] Metzinger, 2025.
[15] Natale, 2022.
[16] Floridi, 2023.
[17] Misselhorn, 2022.
[18] Dennett & Caruso, 2021.
[19] Natale, 2022; Ayad & Plaks, 2025.
[20] Floridi, 2023.

instead requires moving beyond functional accounts of agency and examining the social practices through which agents hold one another accountable. It is precisely to this "normative turn" that the next section is dedicated.

4 THE NORMATIVE TURN: MORAL RESPONSIBILITY AND THE SPACE OF REASONS

The failure of the ingredients approach has left us with little to go by in terms establishing in AI systems may or may not possess internal cognitive faculties that would render them ethical agents and recipients. This, however, motivates the search for another, perhaps more fruitful, approach, that explores the idea that moral agency could be justified not by a specific defining property of the agents, but rather by a social and normative practice of responsibility attribution.

Wilfrid Sellars[21] famously distinguished what is known as the "Space of Reasons" from the physical order of nature. In a famous passage of "Empiricism and the Philosophy of mind" he explains that characterizing a state as a state of knowing means placing it within the "logical space of reasons" and not merely giving it an empirical description. Human beliefs, decisions and actions, while being obviously part of the causal chain of events that occur in the world, are also subject to justification, criticism and evaluation. To take part in the Space of Reasons is therefore to become a potential giver and demander of justifications. As John Mcdowell[22] observes, Sellars endorses a view for which the concept of knowledge belong to a normative context, which also requires the ability to be in a certain position to state knowledge about certain facts with a certain sort of entitlement. On this view, a knowledge claim is not an empirical description of a mental episode, but precisely the placing of that claim in inferential relations to other knowledge claims.

As Robert Brandom has subsequently argued by advancing an inferentialist sematic theory[23], this participation in normative practices involves a reciprocal attribution of commitments and entitlements that has an essential social articulation: there is, in principle someone that asks for justifications, but also someone who can be asked, "because the space of reasons is a normative space, it is a social space"[24]. The genuine reporter of an assertion can tell what follows from it and also what would represent valid evidence for it. Brandom calls this "inferential articulation" that plays a crucial role in reasoning: nothing that hasn't got the ability to distinguish some claims or beliefs as being reasons for others can be a concept user of a "knower". For Brandom, the judgements of humans automatically entails taking responsibility for justifying them to other human agents. The problem around moral responsibility ceases to be an attribute of the single agents and emerges within practices of mutual recognition and accountability. In Brandom's pragmatic understanding of the Space of Reasons, taking part in normative practices means entering in a game of commitments and entitlements where rationality becomes a practical and conversational activity. When we enter in the Space of Reasons, we are taking part in a social practice where our assertions are passible of evaluation, critics and justification requests. A crucial, yet overlooked, feature of this game is that it is inherently coercive, in the sense that undertaking a certain commitment means risking losing one's entitlement is that commitment is unfulfilled. Normative status is not permanent, but an extremely fragile standing that can be easily diminished by criticism or failure to provide adequate justifications. Peter Strawson's[25] account of reactive attitudes further strengthens this view, illustrating how responsibility is embedded within interpersonal practices. Emotions such as resentment, gratitude, indignation and forgiveness constitute the normative framework through which agents hold one another accountable. It is precisely on these premises that the necessity of what I shall define as "Symmetrical Vulnerability" comes into focus. For an individual to genuinely occupy the Space of Reasons, I argue it must not only be able to account for the epistemic value of its responses, but it must also be able to *bleed*, meaning it must be capable of experiencing the loss of its normative standing.

5 SYMMETRICAL VULNERABILITY

Now that the question about the secret ingredient to be a moral agent has subsided to the understanding of moral agency as emerging from a normative practice justified by reason demand and attribution, a legitimate question would be the following: what is then the secret ingredient to participate in practices of accountability?

Maybe readers have already started to question the relevance of the solution proposed by this paper, since it seems to be going in the same direction that was somewhat criticized in the opening. Is then the participation to the practice of accountability something that depends on a secret ingredient that only humans possess? The short answer to this question is no; still, this paper argues that the practice of holding one another responsible is not simply a matter of describing an agent's

[21] Sellars, 1956.
[22] Mcdowell, 2018.
[23] Brandom, 1994, 1995, 2000, 2009.
[24] Brandom, 2009, p.4.
[25] Strawson, 1962.

behaviour. Instead, it involves having the ability to direct demands, criticism, blame and praise towards them.

But if machines cannot be held responsible, what are the consequences for the epistemic status of machine's assertions? Robert Sparrow and Gene Flenady[26] introduce the notion of "testimony gap" to describe the situation where there is nobody to take responsibility for the epistemic claims made by an AI system. Following their interpretation of Brandom's inferentialism, it is in fact only the vouching of a human being that can turn the output of an AI into a judgement. The judgement may then be brought up in the logical Space of Reasons, and the human agent might be asked to answer for it. However, in this case the testimony gap remains as the responsibility for the epistemic value of the outputs is now all placed on the human, and not on the machine.

In addition to this, recent work by Shannon Vallor and Tillman Vierkant[27] provides another crucial point. In their analysis of the problem raised by responsibility gaps, they argue that the issue is not merely the absence of adequate knowledge or control of the agency, but rather what they define as a "vulnerability gap". Increasingly complex scenarios involving the interaction of complex social settings and more or less autonomous AI systems often leave open the dangerous possibility of affected individuals not having an identifiable agent who stands in an appropriate relationship of answerability toward them. At the same time, this phenomenon is also causing this fragmentation in human action and decision networks more prominent, often undermining established practices of responsibility attribution. While Vallor's Vierkant's work primarily addresses the issue from the perspective of the human victims in an asymmetrical accountability relation, my proposed account also wants to focus on the active capacity of the agent who trusts AI systems to lose their normative status and undergo reputational degradation.

The analyses I have just mentioned raise an important question and shift attention away from the ingredients approach toward the relational conditions that are needed to sustain moral responsibility. What matters is not simply whether an agent can act autonomously, but whether they can participate at both ends of an accountability practice. On this view, simply taking part in the order of natural phenomena will not do: moral responsibility involves more than causal contribution in an action or decision-making abilities. Responsibility requires the possibility of being called to account for our judgements and our actions by others and to also be exposed to the consequences of one's commitments.

[26] Sparrow & Flenady, 2025.
[27] Vallor & Vierkant, 2024.

Building on this insight, I propose a notion of Symmetrical Vulnerability, which refers to the reciprocal exposure of participants in the practice of accountability to normative sanction. It is not, how it may seem at first glance, a matter of simple physical fragility or susceptibility to harm. Even though physical vulnerability should also be taken into account, the core concept is rather the one regarding a specifically normative form of vulnerability: the possibility of having one's past actions questioned, challenged and evaluated by others. On this account, the significance given to the symmetry derives from the reciprocal structure of accountability. Moral agents have to be both moral judges and moral receivers on order to participate in the practice; occupying both positions simultaneously is what grants the participation in a community. This reciprocal exposure is what has justified centuries of legal practices in many different declinations, and it explains why responsibility practices play such a crucial role in our daily lives. As Vallor and Vierkant emphasise, practices of moral addressing contribute to the development of trust and solidarity.

6 WHAT WE SHOULD BE WORRYING ABOUT

The account I have so far outlined allows to now reconsider the starting question about artificial moral agency from a radically different perspective. The issue is no longer whether AI systems can be described as functional agents, nor whether future implementation may give rise to systems indistinguishable from humans. The relevant questions is whether artificial systems can occupy both ends of the reciprocal normative relation required by practices of accountability.

This paper suggests that current AI systems certainly seem unable to do so; while they may give the impression of influencing human decisions and behaviour, they cannot meaningfully participate in practices of moral address. Why? The answer is rather simple: they have nothing at stake. Not only they are unfit to make judgements in the logical Space of Reasons, AI systems also cannot genuinely acknowledge blame, be subjected to social shaming, experience the loss of normative standing or assuming responsibility for damaged goods or individuals. Consequently, they can certainly not occupy at least one of the two ends of the accountability practice.

Therefore, while providing a satisfactory and empirically verifiable way to close these gaps falls outside the immediate scope of these paper, the relational approach here presented aims to at least help partly redefine the necessary parameters to address them. What should now in any case be clear is that moral responsibility is not transferable to machines, and that it will only become increasingly difficult to identify who should answer in

case of negative outcomes, biased results or unjustified epistemic claims. At the same time, it seems it would not be fair to hold human operators of manufacturers morally responsible for a judgement to which they were not entitled or a harm that they could not have possibly foreseen. The issue then is not the possibility of the emergence of new ethical agents in the moral community, but rather the risk of not having someone to ask for answers when our trust is proven to be misplaced. The problem about AI ethics has recently precisely become one about how to preserve and redesign practices of accountability under conditions of distributed and technologically mediated action. This requires ensuring that human actors remain appropriately answerable for the systems they design, deploy and allow to penetrate decision making settings. Agency cultivation approaches are precisely aimed at this, reconceiving responsibility practices as compatible with the absence of adequate control or entitlement by the human agents involved.

7 CONCLUSION

The question of whether AI systems qualify, or could potentially qualify in the future, as moral agents has been traditionally approached by trying to identify a set of internal faculties thought to also be constitutive of human moral agency, including consciousness, autonomy, free will and control over the consequences one's action. While these features remain at the centre of the debate around human agency, this paper has suggested that none of them provides a stable enough criterion to resolve the pressing matter of the ethical status of AI systems. The resulting debates frequently become trapped in disagreements concerning the possession of particular capacities, therefore leaving unexplored the possibility that moral responsibility could be grounded elsewhere.

To address this issue, I proposed a shift in perspective. Rather than treating moral agency as a property grounded exclusively in internal characteristics of individual agents, I endorsed the idea that it should instead be considered as a shared normative practice within which responsibility is attributed by being able to ask and be asked for justification. Drawing on Robert Brandom's understanding of the sellarsian notion of the Space of Reasons and on contemporary accounts of responsibility as a practice of answerability and entitlement, I argued that moral agency is inseparable from participation in relations of mutual accountability.

The concept of Symmetrical Vulnerability was introduced to capture this relation. Moral agents are not individuals equipped with a series of capabilities, but rather they are participants in practices of justification, criticism, blame, praise and normative sanction. I argued that such practices presuppose a reciprocal structure: those who hold others accountable must also be vulnerable to being held accountable themselves. Symmetrical vulnerability therefore constitutes a necessary condition for participation in the normative community within which the very same concept of moral responsibility becomes meaningful. This perspective has important implications for contemporary debates: even if AI systems were shown to partly satisfy functional accounts of agency, this wouldn't necessarily imply their status of moral agents. As explained above, the central issue this paper aimed to tackle was whether AI systems can occupy the reciprocal normative position required to be held accountable, and they certainly do not appear able to do so.

The conclusion of this paper shifts instead the attention to another core ethical challenge of the debate around AI systems, which consists in an inevitable consequence of the delegation of morally relevant decisions to AI systems: the challenge of testimony gaps and responsibility gaps[28]. As AI systems become more and more embedded in our everyday lives, the redistribution of human agency in these technologically-mediated scenarios needs to be addressed, and both testimony gaps and responsibility gaps, despite their controversial nature and extent, are generally attributed to a failure of the epistemic and/or the control condition of moral responsibility[29]. Artificial Morality is a central part of the implementation of systems that hold great power over society. The current research on the matter provides empirical evidence on the importance of bridging responsibility gaps from both a psychological and motivational perspective. The more intelligent and autonomous technologies get, the more they become part of our lives and the more intricate the moral problems they encounter will become[30]. This is a critically urgent matter, as it puts the very practice of reciprocal accountability at risk of collapsing. The problem of responsibility gaps does not, however, seem unsolvable: if, instead of fatalistically assuming that artificial moral agents are inevitable, the focus becomes keeping humans on-the-loop in a way that lets them stand in the relation required by the practice of mutual accountability, the issue could at least in principle be contained. After all, as Sparrow and Flenady beautifully phrase, "we can only rely on machines insofar as we trust each other"[31].

[28] Sparrow & Flenady, 2025; Vallor & Vierkant, 2024.
[29] Vallor & Vierkant, 2024.
[30] Misselhorn, 2022.
[31] Sparrow & Flenady, 2025, p.14.

REFERENCES

- Ayad R., Plaks J. E. (2025). Attributions of intent and moral responsibility to AI agents. *Computers in Human Behavior: Artificial Humans*, 3. https://doi.org/10.1016/j.chbah.2024.100107.
- Brandom, R. B. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press.
- Brandom, R. (1995). Knowledge and the Social Articulation of the Space of Reasons [Review of *Knowledge and the Internal*, by J. McDowell]. *Philosophy and Phenomenological Research*, *55*(4), 895–908. https://doi.org/10.2307/2108339
- Brandom, R. B. (2000). *Articulating reasons: An introduction to inferentialism*. Harvard University Press.
- Brandom, R. B. (2009). *Reason in philosophy: Animating ideas*. Belknap.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2(3), 200-219.
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Dennett, D. C. (1984). Elbow room: The varieties of free will worth wanting. MIT Press.
- Dennett, D. C. (1987). *The intentional stance*. The MIT Press.
- Dennett, D. C. (2003). Freedom evolves. Viking Press.
- Dennett, D. C., & Caruso, G. D. (2021). *Just deserts: Debating free will*. Polity Press.
- Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge; New York: Cambridge University Press.
- Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. Minds and Machines, 14(3), 349--379.
- Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. https://doi.org/10.1093/oso/9780198883098.001.0001
- Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, *68*(1), 5-20
- Mcdowell, J. (2018). Sellars and the Space of Reasons. Analysis. Claves de Pensamiento Contemporáneo, 21 (8), pp.1-22. ⟨10.5281/zenodo.2560006⟩.
- Metzinger, T. (2025). Applied ethics: synthetic phenomenology will not go away. Front Sci 3:1702840. doi: 10.3389/fsci.2025.1702840
- Misselhorn, C. (2022). Artificial Moral Agents: Conceptual Issues and Ethical Controversy. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (pp. 31–49). chapter, Cambridge: Cambridge University Press.
- Nagel, T. (1974). What is it like to be a bat?. *The Philosophical Review*, *83*(4), 435-450. DOI: 10.2307/2183914.
- Natale S., (2022). *Macchine ingannevoli. Comunicazione, tecnologia, intelligenza artificiale*. Torino: Einaudi.
- Pereboom, D. (2022). *Free Will*. Cambridge: University Printing House.
- Porter, M. S., (2024). Moral agency in Silico: Exploring Free Will in Large Language Models. arXiv.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- Sellars, W. (1956). Empiricism and the philosophy of mind. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science* (Vol. 1, pp. 253–329). University of Minnesota Press.
- Seth, A. (2025). Conscious artificial intelligence and biological naturalism (Version 1). University of Sussex. https://hdl.handle.net/10779/uos.28649519.v1
- Sparrow, R. & Flenady, G. (2025). The Testimony Gap: Machines and Reasons. *Minds and Machines* 35 (1):1-16.
- Strawson, P. F., (1962). Freedom and Resentment. Proceedings of the British Academy, 48, 187-211.
- Vallor, S., & Vierkant, T. (2024). Find the Gap: AI, Responsible Agency and Vulnerability. *Minds and machines*, *34*(3), 20. https://doi.org/10.1007/s11023-024-09674-0.