



# Moral Responsibility and Machine Agency in Contemporary Debates on Artificial Minds: A Review of the Philosophical Literature

Stefan Mansour<sup>1\*</sup>, Tigran Harkness<sup>2</sup>, Sipho Radovic<sup>3</sup>

<sup>1</sup>Department of Social Sciences, Vistula Center for Social Sciences, Malaysia

<sup>2</sup>Department of Social Sciences, Tagus Research Institute of Social Sciences, Cote d'Ivoire

<sup>3</sup>Department of Social Sciences, Chaliyar Institute for Advanced Social Sciences, Brazil

## Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijpts.2026.v1.i1.pp1-7](https://doi.org/10.62715/ijpts.2026.v1.i1.pp1-7)

Received: 21 April 2026

Accepted: 08 June 2026

Published: 18 August 2026

## Abstract

Machines that learn, plan and act with little human supervision have revived an old philosophical question in a practical form: when such a system causes harm, who, if anyone, is morally responsible. This narrative review surveys the philosophical literature on that question from the early work on machine ethics to recent debates about responsibility gaps. It first sets out the competing accounts of what moral agency requires, contrasting the standard view, which ties agency to mental states and understanding, with functionalist accounts that assess agents at a chosen level of abstraction and with Strawsonian accounts that locate responsibility in interpersonal practices. It then examines the argument that learning systems create a gap in responsibility, the different kinds of gap that later authors have distinguished, and the four main families of response: denying the gap, closing it through meaningful human control, distributing responsibility across collectives, and accepting that some harms will go unanswered. The review finds that much apparent disagreement stems from different senses of responsibility, and that the least developed part of the debate concerns the history of an artificial agent's values and the standing of those affected to demand answers.

**Keywords:** *Moral responsibility, Artificial agency, Responsibility gap, Machine ethics, Meaningful human control, Reactive attitudes*

© 2026 Stefan Mansour et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## 1. Introduction

The question whether a machine could think was posed in its modern form by Turing (1950), who proposed replacing it with a test of conversational behaviour. The question whether a machine could be responsible for what it does is younger and has a different character. It cannot be settled by behaviour alone, because responsibility is a normative status: to call an agent responsible is to say that certain responses to it, such as blame, praise, punishment or a demand for explanation, are fitting. Any answer therefore depends on a theory of what makes those responses fitting.

\* Corresponding author: Stefan Mansour, Department of Social Sciences, Vistula Center for Social Sciences, Malaysia. E-mail: [stefan.mansour@vistula.edu](mailto:stefan.mansour@vistula.edu)

© 2026 Stefan Mansour et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The practical occasion for the debate is the spread of systems whose behaviour is shaped by training on data and not specified line by line. A conventional program does what its authors wrote, and its failures can be traced to a design decision or a coding error. A learning system acquires dispositions that its developers did not choose individually and may be unable to predict in detail. When such a system contributes to a wrongful arrest, a fatal collision or an unlawful strike, the familiar chain from outcome to responsible person seems to weaken.

This review organises the philosophical literature on that problem. It asks what the main accounts of moral agency imply for artificial systems, what the responsibility gap argument claims and how it has been refined, and which responses are currently defended. The aim is to show where the positions actually conflict and where they are answering different questions under a shared vocabulary.

## 2. Scope and Approach

The article is a narrative review of the published philosophical literature. It draws on work in the ethics of technology, moral psychology and the theory of responsibility, selected for its influence on the debate and its usefulness in marking out the main positions. It does not attempt an exhaustive search, and it leaves aside legal liability except where philosophers have used legal categories as models. The related question of whether artificial systems could be moral patients, owed consideration for their own sake, is set aside, although a complete treatment of artificial minds would have to address agency and patiency together.

## 3. Accounts of Moral Agency

Whether a machine can bear responsibility depends first on whether it can be a moral agent at all. Three broad accounts structure the literature. Table 1 summarises them, together with two positions that follow from them, and the subsections below take each in turn.

| Table 1: Main positions on the moral agency of artificial systems |                                                                                       |                                                                      |                                                        |
|-------------------------------------------------------------------|---------------------------------------------------------------------------------------|----------------------------------------------------------------------|--------------------------------------------------------|
| Position                                                          | Core claim                                                                            | Representative sources                                               | Main difficulty                                        |
| Standard view                                                     | Agency needs mental states, intention and understanding; machines lack them           | <a href="#">Johnson, 2006</a> ; <a href="#">Searle, 1980</a>         | Relies on contested claims about machine understanding |
| Functionalism                                                     | Agency is fixed by interactivity, autonomy and adaptability at a level of abstraction | <a href="#">Floridi and Sanders, 2004</a>                            | Separates agency from blameworthiness                  |
| Graded agency                                                     | Machines occupy stages between ethical impact and full agency                         | <a href="#">Moor, 2006</a> ; <a href="#">Wallach and Allen, 2009</a> | Says little about who answers for intermediate stages  |
| Practice-based view                                               | Responsibility depends on fitting reactive attitudes within relationships             | <a href="#">Strawson, 1962</a> ; <a href="#">Shoemaker, 2011</a>     | Human attitudes toward machines may be misplaced       |
| Historical view                                                   | An agent must own the values it acts on; engineered values are not owned              | <a href="#">Hakli and Mäkelä, 2019</a>                               | Humans also acquire values they did not choose         |
| Instrumental view                                                 | Robots are artefacts, and their makers and owners answer for them                     | <a href="#">Bryson, 2010</a>                                         | Strained by systems no one can predict                 |

### 3.1. The standard view

On the standard view an agent is morally responsible only if it acts from intentions, grasps reasons and understands what it is doing. [Johnson \(2006\)](#) applies this view to computers and concludes that they are moral entities without being moral agents. They are moral entities because they are made by people with

purposes and they shape human action; they are not agents because they have no mental states and no intendings of their own. Their behaviour is always a component of human action, and responsibility stays with the humans whose intentions are built into them.

The standard view often draws support from arguments that computation cannot amount to understanding. The best known is the Chinese room argument of [Searle \(1980\)](#), according to which a system that manipulates symbols by formal rules has syntax without semantics. The argument has found a new audience in discussion of large language models, where critics hold that a system trained only on linguistic form has no access to meaning. Opponents appeal to the intentional stance: if treating a system as having beliefs and desires yields reliable prediction, there may be no further fact about whether it really has them ([Dennett, 1987](#)). The dispute matters for responsibility because blame is ordinarily addressed to a being that can grasp why its conduct was wrong.

### **3.2. Functionalist and graded accounts**

[Floridi and Sanders \(2004\)](#) reject the appeal to inner states. They propose that agency be assessed at an explicitly chosen level of abstraction, and that at a suitable level a system counts as an agent if it is interactive, autonomous and adaptable. An agent whose actions can cause moral good or evil is a moral agent. The authors call the result a mind-less morality and accept its most debated consequence: an artificial agent can be accountable, in the sense that it can be the source of a moral action and can be corrected, retrained or switched off, without being responsible in the sense that deserves blame. Critics object that this keeps the word agency while removing what made agency morally interesting.

Graded accounts avoid a single threshold. [Moor \(2006\)](#) distinguishes ethical impact agents, implicit ethical agents whose design constrains them to act safely, explicit ethical agents that represent and reason about ethical considerations, and full ethical agents with consciousness, intentionality and free will. [Wallach and Allen \(2009\)](#) place systems on two axes, autonomy and sensitivity to morally relevant features, and describe a region of functional morality between the operational morality fixed by designers and full moral agency. These schemes are useful for engineering, but they were not written as theories of responsibility, and they leave open who answers for an agent in the middle of the scale.

### **3.3. Responsibility as a practice**

A third tradition begins from [Strawson \(1962\)](#), for whom holding someone responsible consists in being disposed to reactive attitudes such as resentment, gratitude and indignation. These attitudes belong to a participant stance taken toward those with whom ordinary relationships are possible, and they are suspended when the other is seen as an object to be managed. Later writers have divided the concept further. [Shoemaker \(2011\)](#) distinguishes attributability, the sense in which conduct discloses an agent's character, accountability, the sense in which sanctions are fair, and answerability, the sense in which an agent can be asked to justify conduct by citing reasons.

Applied to machines, the practice-based tradition yields a more differentiated question. A system might be answerable in a thin way if it can report the considerations on which an output was based, yet not accountable, because sanction presupposes a being that can suffer or care. The tradition also invites a historical test. [Hakli and Mäkelä \(2019\)](#) argue that robots cannot be morally responsible however sophisticated they become, because the capacities and values from which they act were installed by others, in a way that resembles the manipulation cases used in debates about free will. The reply that human values are also shaped by upbringing shows that the argument needs a principled account of which histories undermine ownership and which do not.

## **4. The Responsibility Gap**

The argument that gives the debate its name was stated by [Matthias \(2004\)](#). Responsibility for the behaviour of a machine has traditionally fallen on its manufacturer or operator, on the ground that they control it. Learning automata weaken that ground. Their rules of operation are altered during use, by the machine itself, and the manufacturer and operator are in principle unable to predict the resulting

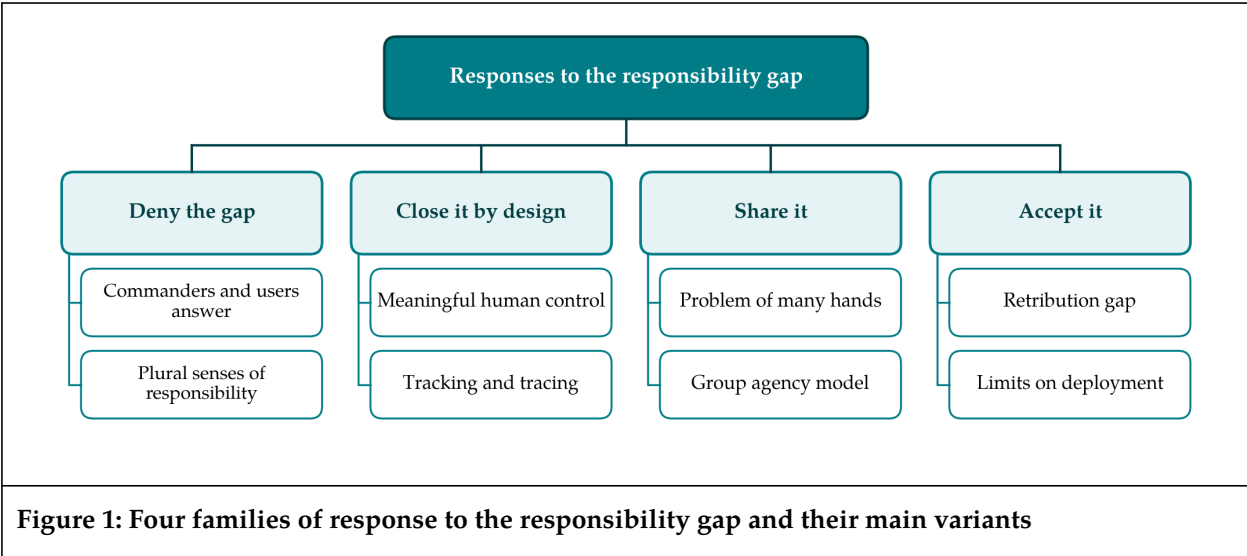
behaviour. If control is a condition of responsibility, they cannot fairly be held responsible, and since the machine is not a suitable bearer either, there is a gap. Sparrow (2007) gave the argument its sharpest application. If an autonomous weapon commits what would be a war crime, neither the programmer, nor the commanding officer, nor the machine can justly be held responsible, and he concludes that deploying such weapons is therefore wrong.

Both authors rely on a control condition of the kind analysed by Fischer and Ravizza (1998), who hold that responsibility requires guidance control: the action must issue from a mechanism that is responsive to reasons and that the agent has made their own. The gap arises when no human exercises such control over the particular outcome. The difficulty of locating responsibility in computerised systems had been noticed earlier. Nissenbaum (1996) identified four barriers to accountability: the problem of many hands, the acceptance of bugs as inevitable, the use of the computer as a scapegoat, and ownership without liability. What learning systems add is a barrier that does not depend on negligence or evasion by anyone.

Later work has shown that the gap is not one thing. Danaher (2016) distinguishes a retribution gap: people have a strong desire to punish when harm is done, robots are not fitting objects of retributive blame, and often no human will be either, so the desire goes unsatisfied or is misdirected. Santoni de Sio and Mecacci (2021) separate four gaps, in culpability, in moral accountability, in public accountability and in active responsibility, the last being the forward-looking duty of designers and users to ensure that systems behave well. The distinctions matter because a measure that closes one gap, such as a compensation fund, may leave the others open.

5. Responses to the Gap

Four families of response can be distinguished, and Figure 1 arranges them. They are not mutually exclusive, and several authors combine elements of more than one.



5.1. Denying the gap

One response holds that existing concepts already cover the cases. Himmelreich (2019) argues against Sparrow that a commander who deploys an autonomous weapon stands to its conduct much as a commander stands to the conduct of subordinates, and that unpredictability in detail does not remove responsibility for deploying a system known to be unpredictable. Tigard (2021) argues more generally that there is no techno-responsibility gap once responsibility is understood pluralistically. Attributability, answerability and accountability can each be secured in some measure, by different parties and by different means, even when no single person satisfies all three. The instrumental view of Bryson (2010) reaches a similar conclusion from another direction: robots are artefacts that people design and own, and presenting them as persons obscures the human decisions for which humans should answer.

### 5.2. Closing the gap by design

A second response accepts that gaps can arise and treats their prevention as a design requirement. [Santoni de Sio and van den Hoven \(2018\)](#) develop the notion of meaningful human control, taken from the debate on autonomous weapons, into a philosophical account built on Fischer and Ravizza. A system is under meaningful human control when it satisfies a tracking condition, by responding to the relevant moral reasons of the relevant humans and to the facts of its environment, and a tracing condition, by allowing its outcomes to be traced to at least one human who understands its capabilities and their own responsibility for its use. On this account control does not require a hand on the lever at each moment. It requires that the whole arrangement of people and machine be built so that responsibility has somewhere to rest.

### 5.3. Sharing responsibility

A third response moves from individuals to collectives. The problem of many hands, described for public officials by [Thompson \(1980\)](#), arises whenever an outcome results from the contributions of many people, none of whom meets the conditions for responsibility for the whole. [Nyholm \(2018\)](#) suggests that the agency of automated systems is best understood as collaborative agency in which humans hold a supervisory role, so that the locus of responsibility is the human party to the collaboration. [List \(2021\)](#) compares artificial systems with group agents such as corporations. On the account of [List and Pettit \(2011\)](#), a group can be fit to be held responsible when it faces morally significant choices, has the understanding needed to assess them, and controls what it does. The comparison supplies a precedent for holding a non-human agent to account and a warning that the conditions are demanding.

### 5.4. Accepting the gap

The final response accepts that some harms caused by autonomous systems will have no responsible author and asks what follows. For Sparrow the consequence is a prohibition in the domain where the stakes are highest. For Danaher it is a need to manage retributive expectations that cannot be met. A relational variant shifts attention from the agent to the people affected. [Coeckelbergh \(2020\)](#) argues that responsibility is owed to someone, and that those subject to automated decisions are entitled to reasons. Explainability then matters as a condition for human agents to answer to those they affect, and a system that cannot be explained should be used with corresponding restraint.

## 6. Open Questions

Three matters remain unsettled. The first is conceptual. Authors who deny the gap and authors who affirm it frequently use different senses of responsibility, and a conclusion about answerability does not settle a question about desert. Progress depends on stating which sense is at issue, as the more recent taxonomies begin to do.

The second concerns history and ownership. If engineered values disqualify a machine from responsibility, an account is needed of why training differs from education, and of whether a system that revises its own dispositions over a long period could come to own them. The literature on manipulation in the free will debate offers materials, but they have been applied to artificial agents only in outline.

The third concerns the standing of those affected. Most of the debate asks who may be blamed, and much less of it asks who is entitled to demand an account and in what form. Systems that produce fluent explanations of their outputs make this pressing, because an explanation that sounds like a reason is not thereby the reason on which the system acted. Whether such outputs can support answerability, or only imitate it, is a question on which the philosophy of responsibility and the technical study of interpretability have so far had little contact.

## 7. Conclusion

The literature reviewed here does not support a simple verdict that machines can or cannot be responsible. It supports a set of conditional claims. If responsibility requires understanding and owned values, current systems are not responsible, and the task is to keep human responsibility intact through

design and institutional arrangement. If responsibility is divided into attributability, answerability and accountability, some of its functions can be met by systems and organisations while others cannot. The responsibility gap is best treated as a family of distinct problems, each requiring its own remedy, and the most promising work is that which specifies in advance which humans are to answer for which aspects of a system's conduct.

## Conflict of interest

To be declared by the author(s) before publication.

## Funding

To be declared by the author(s) before publication.

## References

- Bryson, J.J. (2010). [Robots should be slaves](#). In: Y. Wilks (ed.), *Close Engagements with Artificial Companions*. John Benjamins, Amsterdam.
- Coeckelbergh, M. (2020). [Artificial intelligence, responsibility attribution, and a relational justification of explainability](#). *Science and Engineering Ethics*, 26: 2051-2068.  
<https://doi.org/10.1007/s11948-019-00146-8>
- Danaher, J. (2016). [Robots, law and the retribution gap](#). *Ethics and Information Technology*, 18(4): 299-309.  
<https://doi.org/10.1007/s10676-016-9403-3>
- Dennett, D.C. (1987). [The Intentional Stance](#). MIT Press, Cambridge, MA.
- Fischer, J.M. and Ravizza, M. (1998). [Responsibility and Control: A Theory of Moral Responsibility](#). Cambridge University Press, Cambridge.
- Floridi, L. and Sanders, J.W. (2004). [On the morality of artificial agents](#). *Minds and Machines*, 14(3): 349-379. <https://doi.org/10.1023/B:MIND.0000035461.63578.9d>
- Hakli, R. and Mäkelä, P. (2019). [Moral responsibility of robots and hybrid agents](#). *The Monist*, 102(2): 259-275.
- Himmelreich, J. (2019). [Responsibility for killer robots](#). *Ethical Theory and Moral Practice*, 22: 731-747.
- Johnson, D.G. (2006). [Computer systems: Moral entities but not moral agents](#). *Ethics and Information Technology*, 8(4): 195-204. <https://doi.org/10.1007/s10676-006-9111-5>
- List, C. (2021). [Group agency and artificial intelligence](#). *Philosophy & Technology*, 34: 1213-1242.
- List, C. and Pettit, P. (2011). [Group Agency: The Possibility, Design, and Status of Corporate Agents](#). Oxford University Press, Oxford.
- 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>
- Moor, J.H. (2006). [The nature, importance, and difficulty of machine ethics](#). *IEEE Intelligent Systems*, 21(4): 18-21. <https://doi.org/10.1109/MIS.2006.80>
- Nissenbaum, H. (1996). [Accountability in a computerized society](#). *Science and Engineering Ethics*, 2(1): 25-42. <https://doi.org/10.1007/BF02639315>
- Nyholm, S. (2018). [Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility-loci](#). *Science and Engineering Ethics*, 24(4): 1201-1219.  
<https://doi.org/10.1007/s11948-017-9943-x>
- Santoni de Sio, F. and Mecacci, G. (2021). [Four responsibility gaps with artificial intelligence: Why they matter and how to address them](#). *Philosophy & Technology*, 34: 1057-1084.  
<https://doi.org/10.1007/s13347-021-00450-x>
- Santoni de Sio, F. and van den Hoven, J. (2018). [Meaningful human control over autonomous systems: A philosophical account](#). *Frontiers in Robotics and AI*, 5: 15. <https://doi.org/10.3389/frobt.2018.00015>

- Searle, J.R. (1980). *Minds, brains, and programs*. *Behavioral and Brain Sciences*, 3(3): 417-424.  
<https://doi.org/10.1017/S0140525X00005756>
- Shoemaker, D. (2011). *Attributability, answerability, and accountability: Toward a wider theory of moral responsibility*. *Ethics*, 121(3): 602-632. <https://doi.org/10.1086/659003>
- Sparrow, R. (2007). *Killer robots*. *Journal of Applied Philosophy*, 24(1): 62-77.  
<https://doi.org/10.1111/j.1468-5930.2007.00346.x>
- Strawson, P.F. (1962). *Freedom and resentment*. *Proceedings of the British Academy*, 48.
- Thompson, D.F. (1980). *Moral responsibility of public officials: The problem of many hands*. *American Political Science Review*, 74(4): 905-916. <https://doi.org/10.2307/1954312>
- Tigard, D.W. (2021). *There is no techno-responsibility gap*. *Philosophy & Technology*, 34: 589-607.  
<https://doi.org/10.1007/s13347-020-00414-7>
- Turing, A.M. (1950). *Computing machinery and intelligence*. *Mind*, 59(236): 433-460.  
<https://doi.org/10.1093/mind/LIX.236.433>
- Wallach, W. and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, New York.