AI and Ethics
https://doi.org/10.1007/s43681-025-00685-w

RESEARCH

Uncovering the gap: challenging the agential nature of AI responsibility problems

Joan Llorca Albareda[1]

Received: 28 October 2024 / Accepted: 14 February 2025

Abstract
In this paper, I will argue that the responsibility gap arising from new AI systems is reducible to the problem of many hands and collective agency. Systematic analysis of the agential dimension of AI will lead me to outline a disjunctive between the two problems. Either we reduce individual responsibility gaps to the many hands, or we abandon the individual dimension and accept the possibility of responsible collective agencies. Depending on which conception of AI agency we begin with, the responsibility gap will boil down to one of these two moral problems. Moreover, I will adduce that this conclusion reveals an underlying weakness in AI ethics: the lack of attention to the question of the disciplinary boundaries of AI ethics. This absence has made it difficult to identify the specifics of the responsibility gap arising from new AI systems as compared to the responsibility gaps of other applied ethics. Lastly, I will be concerned with outlining these specific aspects.

Keywords Responsibility gap · Control condition · Agential · Ethical proliferation · Many hands · Collective agency

1 Introduction

In a recent article, Frank and Klincewicz [21] have inquired about the disciplinary boundaries of artificial intelligence (AI) ethics. They argue that many of its philosophical problematics are not specific to this field. However, some indeed are, particularly one stands out above the rest: the responsibility gap. This ethical conflict was first theorized by Matthias [44] in the following terms. New AI systems, linked to the emergence of machine learning, can make decisions and generate results outside of their programming. That is, the programmers do not sufficiently *control* what the machine does, so it is difficult to attribute responsibility to them. Nor can we attribute responsibility to the AI system itself, since it lacks the necessary condition for the attribution of responsibility, i.e., full moral agency. This creates a dilemma: either we forgo the benefits of these systems or accept the resulting responsibility gap [65].

This interpretation of the consequences of the introduction of new AI systems is markedly fatalistic [60].[1] It seems that their implementation leads us, necessarily, to the responsibility gap. For this reason, three major strategies have been developed to oppose the inevitability of the responsibility gap. First, some authors have argued that the concept of responsibility is much richer and more dynamic than this interpretation suggests. It can be understood pluralistically [69, 70] or subject to changing social understandings of the concept [30]. Second, another group of authors have argued that the responsibility gap is inconsistent or of little importance [29, 62]. Responsibility problems arising from technology have always existed and we have tools to counteract them. Third, some authors have offered novel solutions to address the responsibility gap. They acknowledge that it is a real and important problem, but argue that we have means to mitigate the problem and its effects [33, 52].[2]

In this article, I will argue that each of these three strategies has important limitations. The first shows the complexity of

✉ Joan Llorca Albareda
joanllorca@ugr.es

[1] Department of Philosophy I, Faculty of Psychology, University of Granada, Campus of Cartuja, 18011 Granada, Spain

[1] For an interesting critique of traditional approaches to the responsibility gap, particularly applied to Sparrow's [65] contribution to the LAWS (Lethal Autonomous Weapon Systems) debate, see [38].

[2] There are other interesting conceptualizations of the different views on the responsibility gap, for example in Tigard [68], and Mirza-eiGhazi and Stenseke [46].

Published online: 28 February 2025

the concept of responsibility, but does not challenge the fact that the responsibility gap continues to concern a particular notion of responsibility. The second shows, relevantly, the overemphasis that has been placed on this issue, but fails to recognize the importance of responsibility concerns. The third offers good avenues for addressing the responsibility gap, but performs a piecemeal analysis of its foundations and consequences. Given these shortcomings, I will adduce that the responsibility gap arising from new AI systems requires a systematic analysis of the agential dimension. That is, it is presupposed that this ethical problematic arises because of the type of agents these entities are. Two theoretical consequences will follow from this analysis: (i) the responsibility gap is reducible to the problem of many hands and collective agency; (ii) the current treatment of the responsibility gap responds to a lack of interest in analyzing the disciplinary boundaries of AI ethics. Both of these consequences will lead me to ascertain what are the specific aspects of the responsibility gap arising from new AI systems.

This contribution represents a further defense of the views that reduce the AI responsibility gap to the problem of the many hands [31, 75] and that understand it through collective agency [26, 39]. However, it is added that, if we carry out a systematic analysis of the agential dimension, the responsibility gap can be understood as a disjunctive between the problem of many hands and collective agency depending on the type of conception of AI agency from which one starts. Moreover, this article does not only provide an answer to the question of what kind of moral problem the responsibility gap is,[3] but also explains why it has been commonly understood as a new moral problem. The absence of disciplinary analyses in AI ethics, also absent in general terms in the ethics of technology [1], has prevented an understanding of the specific aspects of the responsibility gap produced by new AI systems.

The argument will proceed as follows. First, I will explain what the responsibility gap is in general and why the problem produced by AI can be understood as an agential problem. Then I will analyze why the answers offered so far are insufficient. Second, a systematic agential analysis will be carried out from three approaches to understanding the agency of AI: agential moral responsibility, non-responsible moral agents, and instruments. This analysis will bring me to the conclusion that the responsibility gap boils down to the problem of many hands and collective agency. Finally, I will argue that the responsibility gap highlights both the importance of the proliferation problem in AI ethics and the current underdevelopment of the ethical aspects specifically relevant to this debate. Building on this, I will then develop the ethically relevant aspects of the responsibility gap arising from new AI systems: their internal constitution and external features.

2 Focusing the debate: the agential problem in the responsibility gap

2.1 The general and the specific domains

The responsibility gap is not an ethical problem unique to new AI systems. Responsibility gaps can occur for different reasons and in different contexts. It is therefore important to investigate the nature of responsibility gaps in a general sense. Let us start from a distinction within the concept of responsibility: its *elements* and its *conditions*.[4] On the one hand, Loh [41] has differentiated five elements of responsibility: the responsible subject, the object of responsibility, the patient of responsibility, the regulatory authority of responsibility, and the normative criteria that establishes under which conditions the subject is responsible. She gives the following example:

> [A] thief (the individual subject) is responsible for a stolen book (the retrospective object), or, better, the theft (a sequence of actions that have already occurred) to the judge (the official authority) and towards the owner of the book (the official addressee) under the conditions of the criminal code (the normative criteria that define a legal or criminal responsibility). (p. 2)

I add to these five a new one: the circumstances. The intentions of the agent, the way in which the action took place or her response to the patient constitute an important aspect of responsibility.[5] Not all elements have been of equal importance in philosophical debates on responsibility.

[3] See Oimann [55] for a summary of the stances advocating that the responsibility gap is not a new moral problem. We find some recent defenses in the literature [17].

[4] There are other dimensions of responsibility that I do not consider. Nyholm [53, 54] has pointed out that responsibility has an evaluative (negative or positive) and a projective (retrospective or prospective) aspect. For an interesting application of these ideas, see [15].

[5] This framework proposed by Loh is particularly interesting for this research. I want to elucidate which of these elements is implicated in the AI responsibility gap. As we shall see later, the central issue is that we find agents, potential subjects of responsibility, who do not meet the control condition: neither can they be controlled by human beings nor do they have the moral capacities to exercise sufficient self-control to be responsible. Beyond who suffers the harm (the patient), the action that leads to the harm (the object), the authority to which the harm is responsible (be it a judge, an individual or society) or the normative criteria of responsibility (legal or moral), the responsibility gap focuses mainly on the subjects of responsibility. I have added circumstances as an element because it has a relevant weight in the attributions of responsibility, as we will see later when developing the possible excuses to be held responsible. Although circumstances may

Coeckelbergh, for example, has stressed the need to give more importance to the patient of responsibility [9] and to other objects of responsibility [11].[6]

On the other hand, the literature has identified two main conditions of responsibility: *control* and *epistemic* [18, 58]. The responsible agent must have sufficient control over both herself and her circumstances for us to attribute responsibility to her;[7] and the agent must also have some knowledge of the consequences of the action she is about to take and the circumstances in which she finds herself.[8] Without both of these conditions being met, we can hardly attribute responsibility to the agent.

Matthias [44] has centered the problem of the AI responsibility gap around the control condition: programmers cannot sufficiently control what the machine does, so they cannot be attributed responsibility for the machine's actions and decisions. However, the absence of control need not necessarily produce a responsibility gap. Tigard [68] has categorized two types of excuses in situations where we seek to attribute responsibility for the production of harm according to the control we have over two elements of responsibility. If I do not have sufficient control over the *circumstances*, producing a harm that I am forced to do, I can be excused from responsibility (either by coercion, absence of alternatives, etc.). Here a gap would not arise, but rather the attribution of responsibility would be eliminated. The same would be true if, in attempting to assign responsibility for a harm, we were to realize that this harm was produced by an *agent* who does not have sufficient control over herself. If we realize that it is a child or an animal that has harmed us, we will excuse it. This shows that there are many harms that do not involve blame [29]. It may not be appropriate to attribute responsibility if those who directly or indirectly produce the harm are protected by both types of excuses.

So, what does the responsibility gap consist of? Köhler et al. [34] and Tigard [71] have defined it as a normative mismatch: (i) it is appropriate to hold someone responsible; (ii) there are no candidates that can be held responsible. In the above cases, condition (i) is not met: given the two types of excuses, it is no longer appropriate to hold someone responsible. The AI responsibility gap centers on the agential question, since, due to the autonomy gained by the new AI systems (the kind of agents they are), responsibilities can no longer be attributed to either the programmers or the AI entities themselves [35, 68]. This seems contradictory: if an entity does not have sufficient control over one's own agency, AI case, then it is a good excuse for responsibility not being appropriate; but if responsibility is not appropriate, then it is false that AI systems produce responsibility gaps. This apparent contradiction stems, as I will show shortly, in that no systematic analysis of the agential issue in the responsibility gap has been done so far. If responsibility gaps occur, it must be clarified why the specific type of AI agency produces them and whether other types of agency, other than that of an adult human being, also produce responsibility gaps and cannot be excused.

2.2 Pursued strategies and approach to the systematic analysis of the agential dimension

However, systematic analysis of the agential question has not been the usual response to the responsibility gap arising from new AI systems. As I showed in the introduction, three different strategies have been put forward to account for this problem. In this subsection, I will argue that the weaknesses of each of these strategies lead us to the systematic analysis of the agential question.

First, it has been argued that the responsibility gap takes a very limited notion of responsibility. It is a much richer concept with several dimensions. The main representative of this position is Tigard [68, 69]. Two features of his stance are relevant to the discussion. On the one hand, Tigard holds what he calls the processual view of responsibility [69], which is heir to Strawson's [66] philosophy. The latter understands responsibility as a set of reactive attitudes linked to social beliefs and practices. We react to the actions of others in a certain manner, holding them responsible according to the ways in which we understand human interactions. In this sense, responsibility ceases to be an agent-related property and focuses on reactions and social modes of understanding responsibility [69, 70].[9] The consequence of this is that practices of responsibility are subject to change according to social beliefs and modes of reacting to others. On the other hand, this processual view

be included in the normative criteria, we can consider that they have importance by themselves.

[6] With respect to the object of responsibility, Coeckelbergh emphasizes, given the non-neutrality of technologies, a narrative dimension of responsibility.

[7] It should be clarified that the control condition need not imply alternative courses of action. This is an important issue in the free will debates that I cannot deal with. For a development of the actual-sequence notion of guidance control, see [18].

[8] Generalized ignorance or absence of specific knowledge can also be understood from indirect responsibility. We will develop this issue further below. For a comprehensive discussion, see [81].

[9] It is worth pointing out a recent paper by Oinmann and Tollon [56] on how metaphysical conceptions of responsibility affect the AI responsibility gap debate. They argue that most of the disagreement in this debate stems from opposing conceptions of responsibility: either it is understood from the property-view (what they call response independence or being responsible) or from the processual view (what they call response dependence or holding responsible). To some extent, they make a similar move to the one I undertake in this article, that is, to analyze systematically how a significant moral concept, in my case agency, affects this debate.

of responsibility can be understood in a pluralistic manner. Watson [77] and Shoemaker [64] have taken Strawson's processual view one step further: social dynamics are so complex that these interactive processes accommodate different types of responsibility. It is not the same for us to blame an individual for an action (accountability) as for a character trait reflective of the agent's values (attributability) or for not having provided appropriate explanations for a harm (answerability).[10] Given both of these features of responsibility, Tigard argues against gap fatalism: there remain dimensions of responsibility that AI fits into and we can socially adapt to take them into account.

This argument has considerable merit and shows that responsibility is a more complex issue than is expressed in the gap. However, it does not answer the main challenge of the responsibility gap for two reasons. On the one hand, the Strawsonian view has a strong limitation. As McKenna [45] has argued, social reactions and beliefs must be appropriate, that is, they must have an adequate conception of responsible agency and the relevant circumstances. Not everything goes in social practices. Therefore, we must ask ourselves what kind of agent can be held responsible. On the other hand, different types of responsibility require different responsible agencies. While we can imagine an AI that has ethical programming that allows it to respond appropriately [36] or character in a motivationally thin sense [50], it is difficult to accept that an AI can be accountable, as Tigard himself acknowledges, because full moral agency is required for this kind of responsibility.[11] And accountability is an important practice of human societies as some authors have pointed out [14]. Thus, the responsibility gap seems to refer to a gap in accountability [43, 51] and Tigard does not offer solutions in the face of it.

Secondly, it is argued that the responsibility gap is not appropriate. As I showed in the previous subsection, it seems to be a contradiction in terms. Both Hindriks and Veluwenkamp [29] and Simpson and Müller [62] are of this opinion. The former argue that the responsibility gap is conceptually incoherent and the latter that this problem, already present in many other artifacts, is not of much importance. In both cases, the focus of the problem is shifted to the concept of risk: the fundamental issue is to establish socially acceptable levels of risk in the implementation of new AI systems. Simpson and Müller [62] pose the example of bridge construction: while the bridge may collapse and injure many people without the engineer's intentions being malicious, what is important is that the engineer designs a bridge that is robust and reliable enough to be used. We are willing to assume, as with other technologies, infrequent malfunctions in exchange for ample benefits.

We can identify two problems in this position, one factual and the other normative. On the one hand, as previously pointed out, responsibility practices are relevant elements of human societies. There are psychological causes that account for their importance. Danaher [14], for example, focuses on the retributive aspect of responsibility and its rootedness in human psychology. Thus, it does not seem to be so simple to avoid responsibility issues. On the other hand, the desirability of this position is also in question. The lack of consideration for responsibility may discourage the formation of virtuous agencies or encourage the perpetration of harms that will have no consequences for those who produce them [79]. It can also be argued that responsibility is a value worth defending in its own right.[12]

Third, it is argued that we have conceptual means to bridge the gap. The most promising avenues have combined two approaches to responsibility: indirect responsibility and joint responsibility. The first derives from the so-called *tracing condition* [18], that is, we not only hold others responsible for their direct actions, but we have to look at the effects these actions produce over time and on other entities.[13] The second understands that chains of responsibility usually

[10] Tigard [68] defines these three types of responsibility as follows: "to attribute an action to someone is to think it reflects the underlying cares or commitments, whereas to hold one to account is to engage in more overt forms of blame, like directed anger. Aside from these two sorts of responses, however, we might also demand answers, particularly for the harms that befall us. Answerability, on Shoemaker's account, is a process by which we call upon others to provide explanations, in order to evaluate their judgment and understand their reasons for action" (p. 599).

[11] Tollon [72], by contrast, has argued that AI cannot be responsible in terms of attributability and answerability either. Character demands a kind of moral agency that these systems do not possess, and answerability cannot be reduced to explanations, but must be understood primarily as justifications.

[12] A potential counterargument might be the following. These criticisms do not address the real problem of the responsibility gap: it is an inconsistent moral problem and, therefore, we should reject it. Hindriks and Veluwenkamp [29], for example, raise a similar inconsistency to the one I discussed in the previous subsection: "the thing to note is that, if blame is appropriate, then there is reason to attribute it. And if there is reason to ascribe blame, it must be possible to do so" (p. 4). If AI is not an appropriate agent to be blamed, then we cannot blame it. This leads them to reject the concept of responsibility and embrace the concepts of control and risk. However, this does not solve the source of the problem. We are unwilling to accept the problems of AI responsibility because we care about our responsibility practices. If new AI systems erode these practices, then they are damaging something we hold dear. We can insist that the responsibility gap is conceptually inconsistent, but we will still fail to resolve the claims of those people who seek to attribute responsibility. My thesis, defended in Sect. 3, is that if we dive into the initial inconsistency of the responsibility gap we will find important and conceptually consistent moral problems: collective agency and many hands. I am grateful to an anonymous reviewer for suggesting this counterargument.

[13] Here we can distinguish between intrapersonal and interpersonal responsibility. In the first case, the example of the drunkard is very illustrative [81]: the drunkard, although she has no control over herself, did have control when she decided to drink in large quantities. In

occur [49] and that some agents take responsibility for others according to each other's roles and capabilities. Nyholm [52] and Köhler [33], as we will show below, develop arguments representative of this line.

This strategy has many strengths, as it emphasizes the importance of analyzing the agential dimension of AI, how it relates to other human agencies and how responsibilities are distributed. However, none of its representatives has carried out a systematic and comprehensive analysis of the agential problems of the responsibility gap. A certain notion of joint agency and associated indirect responsibilities is often advocated, but there is no analysis of how the agential dimension found the debate on the responsibility gap of new AI systems. This is what I will do in the next section, in which I will incorporate the conclusions of this one: the need for a systematic analysis of the agential dimension in the responsibility gap, understood the latter as an accountability gap.[14]

3 Agency and the responsibility gap

If new AI systems generate responsibility gaps, it is essential to delve into what are the agential characteristics that lead to this problem and how they differ from other types of agency. As we have seen, the responses to the responsibility gap have been different, but none have attempted to seek their foundations in a systematic analysis of AI agency.

This is not to imply that there have not been many analyses of the type of moral agency that AI possesses. We do find an extensive literature on this topic [3, 12, 25, 28, 42, 63]. However, discussions on the responsibility gap have not sufficiently emphasized the connection between responsibility attributions and the type of agency of AI.[15] If the responsibility gap, as we have seen above, has been understood as an agential problem (the type of agents these entities are), then it matters and very much for the responsibility gap debate how we understand AI agency. In what follows, I will present three types of conceptions of AI agency based on their relationship to the concept of responsibility. This classification pivots on two variables: whether AI is a moral agent and whether AI can be responsible. These typologies will show us that no conception of AI agency justifies understanding the responsibility gap as a new moral problem.

In this sense, AI agency can be understood in three ways: agential moral responsibility, moral agents without responsibility, and instruments that do not possess moral agency. The first view holds that both human and artificial agency can be held accountable. As we will develop later, this equivalence can be understood in two ways: (1) AI possesses full moral agency or (2) the conception of moral agency shifts to incorporate AI responsibility.

The second has been argued by authors such as Moor [47] and Floridi and Sanders [20], who have shown that AI can be understood as a moral agent without responsibility in a functionalist sense: although they do not possess consciousness and other properties fundamental to moral agency, they do perform, in a minimal sense, all the functions associated with moral agency. For Floridi and Sanders these are four: (i) interactivity, (ii) autonomy, (iii) adaptability and (iv) moral impact. New AI systems can interact with the environment and change their internal states based on stimuli (i); they can change their internal states according to their own rules, without human intervention (ii); they can modify their own rules in a self-governed way, without simply deriving this from environmental stimuli (iii); and their actions and decisions have morally relevant effects (iv). These authors consider it very different who is the moral source of certain acts and decisions, and the evaluation or prescription of who is responsible for those acts and decisions and whether this evaluation coincides with the moral source.

The third understands that AI is not substantially different from other technical artifacts and that the treatment and consideration we should have of it should not differ from other instruments. AI is not a moral agent and all considerations about the responsibility for its actions should be attributed to its programmers, marketers and users. In what follows, I will develop the problems of responsibility linked to these three types of agency.

3.1 Agential moral responsibility

AI can be understood as a moral agent with responsibility in two senses. On the one hand, it is argued that human and AI agency are indistinguishable, as both are or can be full moral agents. That is, both possess or can possess the properties necessary for full moral agency, such as conscience or practical rationality [28]. This position is framed within the futurist positions advocating superintelligence [5]. In the future, AI will be as or more intelligent than humans. On the other hand, a radical innovation in the concept of agency is proposed. AI should not be understood as an individual

the second, the indirect effects extend to other entities, be they instruments or people, conditioning and producing negative consequences.

[14] From now on, I will refer to responsibility strictly as accountability.

[15] Some authors have incorporated agential considerations in their analyses of the AI responsibility gap. For example, Coeckelbergh [9] deals with who or what is the agent of responsibility and discusses the many hands problem, or Vallor and Vierkant [75] argue that the conditions that must be met in the classical definition of moral agency do not conform to the cognitive biases found in humans. However, none of them discuss the responsibility gap debate from the different conceptions of AI agency and how each of them affects the ways in which we assign responsibility. I am grateful to an anonymous reviewer for encouraging me to clarify the relationship between AI, agency and responsibility in this debate.

agency, but in a way that resembles a collective agency [19, 26]. Like other artificial entities such as corporations [23, 39], AI is the product of multiple human individuals, and its actions and decisions are subject to mechanisms and workings that go beyond particular human agencies.

The first sense, however, is not particularly relevant to the responsibility gap debate. It can be argued that AI currently lacks such agential capabilities and the plausibility of these futuristic perspectives is seriously questioned [73]. However, the most important reason lies in the fact that it would eliminate the responsibility gap. Recall that this is founded on the fact that neither the programmers nor the AI can be held responsible for the latter's actions. On the contrary, if the AI were a full moral agent, it could be held accountable for its actions. The moral challenges would be others: it would be clear to whom we would assign responsibility.

The second sense gives a suggestive solution to the question. While it is true that no responsibility can be attributed to any particular programmer or to AI in an individual sense, new AI systems can be understood to be collective agents. Discussions around collective agency and the possibility that they are responsible agents have been of particular importance in the analysis of corporations [40]. Pettit [57] has argued that corporations meet the three necessary and sufficient conditions for the attribution of responsibility. First, they are autonomous agents. They are agents because they are capable of forming beliefs and desires and acting on the basis of these. And they are autonomous because these beliefs and desires are not reducible to the intentions of their members, but possess constitutions with well-established purposes and rules that need not reflect the majority interest of their parties. Second, corporations can make judgments about the relative value of relevant options in a decision. While groups do not possess full evaluative capabilities and require their members to make such judgments, individual judgments are made in settings constrained by the corporation's rules and procedures and on the basis of group purposes. Third, corporations are sensitive to the reasons provided by value judgments. This is the most problematic condition and the key aspect in the attribution of responsibility. It seems that those who are sensitive to reasons and end up controlling the decisions relevant to the group are individuals. Everything a group does would be done either by a single individual or by a set of individuals. Pettit believes that this is not the case for the following reason: both the collective and the individual agencies have control over the action, since the group has control through the instructions and rules to be followed by the individuals and the individuals have ultimate control over the concrete action that is performed. The fact that there are different orders of control does not mean that one is reducible to the other.[16]

Gunkel [23], Hanson [26] and List [39] have shown that these arguments apply equally to new AI systems. They possess their own beliefs and desires, and these are autonomous, in the sense that they are not reducible to what is dictated by their programmers (precisely what causes the responsibility gap).[17] They can also evaluate different options by means of the patterns obtained through their training, although their evaluations come originally from the evaluative capabilities of programmers. And, finally, the control regimes are reversed: the programmers set the training instructions by which the rules and patterns emerge, and the systems make decisions according to the rationales derived from these patterns and instructions.[18]

Therefore, new AI systems can be understood as moral agents with responsibility from the notion of collective agency. From this conception, the responsibility gap is diluted: we can indeed assign responsibility to AI.

3.2 Moral agency without responsibility

The central problem of the responsibility gap seems to be centered on this type of agency: artificial entities that possess an agential level that prevents their actions and decisions from being controlled by programmers, marketers and users; whose actions have a morally relevant impact; but,

[16] Pettit [57] presents both levels of control in the following manner: "Things may be perfectly analogous in the case of the group agent, except that the mode of control is different. The group may control in a reason-sensitive way for the performance of a certain action by some members, maybe these or maybe those. It will do this, by maintaining a constitution for the formation and enactment of its attitudes, arranging things so that some individual or individuals are identified as the agents to perform a required task, and other individuals are identified as agents to ensure that should the performers fail, there will be others to take their place as backups. Consistently with this group level control, however, those who enact the required performance will also control in a reason-sensitive way for what is done; they will control for the fact that it is they and not others who actually carry it out" (p. 192).

[17] List [39] notes that the cause or origin of AI systems does not determine its moral agency and responsibility: "Again, the objection misses a key point, namely that, no matter how AI systems have been brought into existence, systems above a certain threshold of autonomy constitute new loci of agency, distinct from the agency of any human designers, owners, and operators. In fact, such systems can arguably become even more autonomous than group agents" (p. 1225).

[18] The understanding of new AI systems as collective agencies can be rejected for several reasons. On the one hand, it may be considered that the conditions put forward by Pettit are not sufficient or that, despite being sufficient, his arguments are not adequate [25]. On the other hand, it can be argued that the collective agency of a corporation is different from that of an AI. Courtenage [13], for example, argues that AI is affected, in this approach, by the problem of covert manipulation: its desires and beliefs are predetermined by its programmers. Corporations, by contrast, are not, because of the supervention of human agencies in their operation.

despite their high agential level, they cannot be held responsible. The key aspect of this perspective lies in the analogy with other agencies. We seem to find cases of minimal agencies, as in the case of children and working animals, that comply with the conditions of functional moral agency, have a moral impact and in which the excuses discussed above have no place. On the contrary, the actions of minimal agencies can be understood from the matrix of indirect responsibility and joint responsibility.[19]

Two approaches have accounted for this analogy: *collaborative responsibility* and *instrumental responsibility*. On the one hand, collaborative responsibility is based on combined actions between agents. Both agents take sides in the implementation of an action and its consequences. The responsibility gap assumes that all agents that are part of a collaborative chain should be able to be held responsible. Nyholm [52] argues that this need not be the case. Within chains of responsibility there are agents that cannot be held responsible and others that can. A full moral agent can push or impel another entity with a lower agential level to perform a task. Like a child performing an action impelled by her parents, AI systems perform tasks inspired, even if not in a fixed and univocal way, by a set of human actors. Studying the type of agency of the child leads us, in Nyholm's view, to see how her actions find behind them a locus of responsibility in the parents. The same is true for AI. If there is a malicious intention on the part of a moral agent that pushes an entity with a lower agential level to perform an action with harmful consequences, the degree of autonomy of the latter, although agentially relevant, does not prevent the full moral agent from being held responsible [35].

On the other hand, it can be argued that the type of relationship we have with AI is not collaborative, since for true collaboration both agents must have common plans and be intentionally directed towards the same goal [33]. The use of entities with minimal agency, as is the case with the use of animals to perform certain tasks, does not exempt its moral connotations from a similar valuation as one would have with any other instrument. It is an instrument that can produce effects that were not fully foreseen, but, nevertheless, it was known, when it was chosen to use it, that it could produce negative consequences. A clear example is that of the drunkard [81]. The drunkard cannot take complete responsibility for herself and many of the actions she performs she would not do if she were sober. Nevertheless, the individual is still indirectly responsible: she chose to drink in large quantities knowing that she would not be able to exercise full self-control while drunk. Those who develop, market, and use the new AI systems may not fully control what they do, but they use them for certain purposes knowing they may produce unforeseen consequences.

De Jong [16] has stressed that Nyholm's approach has an important limitation. While it is possible to understand new AI systems from collaborative responsibility, the attribution of responsibilities would still be complicated. Those human agencies charged with overseeing and instructing AI systems are very large in number and are part of very extensive chains of programmers, developers, marketers, and users. In this sense, it would not be easy to determine who is responsible. The same can be said of Köhler's position. New AI systems are instruments that are designed, built and used by many agents. Therefore, it remains unclear who should be held responsible for the actions of these systems.

This leads to an interesting conclusion. The AI gap has been understood as an agential problem: because of the type of agents they are, we cannot assign responsibilities either to AI systems or to those who program and use them. In the previous subsection, we have already shown that, if we understand new AI systems as collective agents, we can hold them responsible for their actions and decisions. If we understand them as moral agents without responsibility, we have avenues for attributing responsibility as well. There are other minimal agencies that would fall within this conception and for which we have means of assigning responsibility: indirect responsibility and joint responsibility. But, ultimately, these avenues end up revealing a responsibility gap of another nature, that is, the many hands problem. There are so many agencies involved in building, deploying and using new AI systems that it is very difficult to assign responsibilities. I will develop this problem further in the next section.

3.3 Neither moral nor responsible agency

Under this characterization, new AI systems would not differ from other instruments: they would be artifacts programmed, developed and used by morally responsible individuals who could be accountable. Here again, it would be a matter of instrumental responsibility, but without considering the possibility that AI could be a morally relevant autonomous agent. The problem would therefore lie in finding those who have caused the harm and can be blamed for it.

Here appears again the problem of many hands: when certain actions and/or instruments are produced in environments where several individuals are involved, it becomes very difficult to attribute responsibilities. There is a tension between individual and collective responsibility [74]: we cannot impute responsibility to a collective and, due to the large number of individuals who have participated in the action, it is not known how to attribute individual responsibility. Two reasons may preclude the possibility of

[19] See my critique of the third strategy against the fatalistic interpretation of the responsibility gap in Sect. 2.2. The concepts of indirect responsibility and joint responsibility were introduced there.

attributing responsibility to a collective: (i) some authors argue that collectives cannot be responsible, only human individuals are [25]; (ii) the outcomes of the action are not the result of intentional coordination between different individuals [33].

Nissenbaum [51] has shown how the problem of the many hands constitutes one of the great challenges to accountability in computerized societies. She offers four reasons. First, most computer systems are elaborated in organizational settings, which implies that many individuals participate in their elaboration. Second, computers are not monolithic entities, but have many modules and segments. Third, the different software levels of the same system make it difficult to establish which level is causing the problem. Fourth, the close link between the program and the hardware often makes it difficult to establish in which of them the cause is to be found. All four apply to algorithmic systems: they are collaboratively designed, modular, multi-layered, and intertwined with hardware (whether robotic or otherwise), making it difficult to pinpoint the origin of problems.

The last three reasons provided by Nissenbaum shed light on what Coeckelbergh [9] has called the problem of many things: interconnected elements of hardware and software that contribute causally to the action and that may have different degrees of agency. This also complicates the assignment of responsibility, since we do not know which part of the hardware and software has produced a harm and, therefore, it is more difficult to tell which individuals are responsible. In this sense, when I refer to the problem of many hands, I will also be including the problem of many things.

In this regard, the different people who are part of the programming, marketing and utilization process; and the complexity of parts and levels of which algorithmic systems are composed make it difficult to assign responsibility. A responsibility gap arises due to the many hands problem.

3.4 The responsibility gap disjunctive

Systematic analysis of the agential dimension of the responsibility gap raises the following disjunctive. Across the last two conceptions of AI agency, the responsibility gap is not attributable to the distinct properties of the AI agents themselves, but rather to the diffusion of responsibility inherent to the many hands problem. The joint and indirect responsibility arguments show that it is possible to hold us accountable for unaccountable moral agents that we do not sufficiently control; but they are unable to provide an explanation of the ways in which to assign accountability to the various actors present in the programming, commercialization, and use of these systems. This is also the case if we understand that AI systems do not possess any extraordinary agential capacity that differentiates it from other instruments. There is a responsibility gap (someone must be held accountable, but we cannot find any candidate), but this gap depends on the multiplicity of agents, and the complexity of these systems and joint action.

At the same time, I have shown that new AI systems can be understood as collective and responsible agents. This answer shows that, if we accept collective agency, a new challenge to the responsibility gap arises: responsibility should not be sought in particular individuals who are part of the process, but in the coordinated whole in charge of its programming, elaboration and/or use. Thus, either we shift the focus of responsibility to the collective level or, at the individual level, the responsibility gap is reduced to the many hands problem. In neither case, however, can the responsibility gap be derived, fatalistically, from the particular type of individual agencies that these entities are.

4 Theoretical implications

4.1 The underlying problem: ethical proliferation and the lack of boundaries

Let us return to the introduction of the article. Frank and Klincewicz [21] argue that, contrary to many debates in AI ethics, the responsibility gap reveals a very particular problem of this discipline. This conception goes hand in hand with the idea that the kind of individual agency that new AI systems are produces this problem. However, I have shown that the responsibility gap in this type of systems is reducible to a disjunctive: either we understand them as collectively responsible agencies or, at the individual level, it is limited to the many hands problem. Thus, it is not a problem specific to AI ethics and, if it is to be properly addressed, we must deal at a more general level with collective agency and the many hands problem.

The latter has been extensively dealt with in the political sphere. Political decisions often involve a large number of actors in complex institutional settings, endowed with multiple parts and levels. This makes it difficult to know who is responsible for a decision or a harmful state of affairs. This has not prevented attempts to find solutions. Thompson [67] posits that while it is difficult to find the person responsible, it is not impossible. We must properly delimit what kind of excuses are acceptable at the political level and establish legal and social mechanisms to pursue and determine who is responsible. These solutions do not seem alien to AI: they are measures and approaches that converge with the responsibility gap produced by new AI systems. The same is true with collective agency. At the economic level, corporations have been recognized as legal persons, subject to duties and rights that transcend the individuals that compose them [23, 26]. The legal treatment and consideration they receive may provide relevant clues to understand in what sense an AI can be a collective agent [7].

This brings us to the first theoretical consequence: the responsibility gap of new AI systems points to the scant attention devoted to the question of the disciplinary boundaries of AI ethics. What issues does it address? How should it address them? How does it differ from other applied ethics? These are all important questions whose answers greatly condition the appropriateness of approaches to the issues of AI ethics. Some articles have engaged in a meta-reflection on the ethics of AI, i.e., the ethical shortcomings of current approaches and practices in the discipline have been raised. It has been criticized that the principles and rules of regulations subscribed to by public and private bodies have hardly any practical force and move at a high level of abstraction [24, 48]. Heilinger [27] goes a step further and shows, in a much more comprehensive sense, that the concepts, ideas and procedures used in the theory and practice of AI ethics have major limitations.

However, these critiques do not address the question of the disciplinary boundaries of AI ethics. Although they articulate shortcomings within the discipline and offer potential solutions, their analysis is incomplete, as it fails to address the foundational misunderstandings concerning the discipline's legitimate scope and objectives. The absence of such general approaches has led to, as Sætra and Danaher [59] point out, the *problem of ethical proliferation*: since no clear disciplinary boundaries are drawn, ethical problems are discussed that have been dealt with at a more general level or in other ethical disciplines and that do not differ substantially. Therefore, the same object is discussed without taking into account the advances that the same type of discussions has had in other spheres.

The responsibility gap of new AI systems, like other issues in AI ethics, has not regularly and systematically drawn on other discussions of responsibility gaps in other disciplines. This makes it insufficiently linked to general theories of responsibility, of collective agency in the economic realm and of the many hands problem in the political sphere. If the problems of AI ethics were put into dialogue with other more general disciplines and with other applied ethics, the truly specifics of the responsibility gap of new AI systems could be obtained.

This is what Llorca Albareda and Rueda [1] have criticized Sætra and Danaher for: ethical proliferation does not mean that the problematics of AI ethics can be subsumed into the philosophy of technology or other applied ethics, but that we must distinguish those aspects that are general, those that are shared with other applied ethics, and those that are truly specific to AI ethics. This leads us to the second theoretical consequence, which we will develop in the next subsection. That is, if the problem of the responsibility gap of new AI systems is reducible to the disjunctive between collective agency and the many hands, and both problems are well-developed in other disciplines; it must be shown what is specific to collective agency and the many hands in AI ethics.

4.2 What is specific of the responsibility gap arising from new AI systems?

Let us start with an important distinction in philosophy of technology [2, 80]. In analyzing the evaluative dimension of technologies, two types can be identified: internal and social. The first refers to those material consequences or conditions that derive from the internal constitution of a given technology. That is, due to the very technical functioning of an artifact, certain effects arise or certain conditions are created, beyond the social and political context in which it is inserted.[20] The second refers to the effect that certain technologies have given certain social configurations. If such a social configuration did not exist, the effects would not be produced. Winner explains this idea with the example of Robert Moses' bridge: a low-lying bridge is built on the way to the beach that prevents certain vehicles from passing under it. Since the African-American population lacks the financial resources to own a private vehicle, they can only travel by bus to the beach. But the bus cannot pass under the bridge. This does not mean that a low-lying bridge necessarily harms, by its internal constitution, the African-American population, but rather that, given a certain social configuration, the bridge harms the African-American population.

This distinction illuminates two relevant aspects of the specifics of the responsibility gap arising from new AI systems. On the one hand, these systems generate new and more complex objects of responsibility. Algorithms introduce concerns about its direct effects on certain populations and the conditions they create. On the other hand, the social distribution and use of new AI systems pose relevant risks. I will refer in particular to the problems of quantity and ubiquity. I will also point out the importance of our social conception of technology to our understanding of responsibility.

Regarding the first aspect, new AI systems pose two new types of objects of responsibility: large databases that may produce biases and social inequalities; and algorithmic criteria by which the system makes certain decisions. de Bruin and Floridi [6] and Gebru et al. [22] have stressed the

[20] This does not mean that these consequences and/or conditions do not unfold differently in different contexts, but rather that, as we will see below, they are not caused by the context in which they are inserted but by the internal constitution of the technology. Winner [80] gives the example of nuclear power plants, which require a centralist organizational model.

importance of properly dealing with the data by which AI systems are fed and the effects they can have on the whole process of programming, marketing and use. This increases the objects of responsibility: not only malicious actions that introduce biases into databases are [35], but also omissions with regard to properly treating the data of an AI system.

For his part, Kitchin [32] has highlighted the problems produced by the ways in which AI systems are trained. Not only can they be governed by morally harmful patterns and criteria, but, by the very nature of these systems, they are so opaque that they prevent these criteria from being known. Added to all this, both types of responsibility objects produce, in addition to direct consequences, affordances and conditions that affect certain groups of people [43]. Gender or racial biases, for example, can limit the access of groups to certain domains and/or prevent the individuals that compose them from being able to perform adequately in them. Therefore, the increase in objects of responsibility aggravates the responsibility gap: as there are more tasks to be responsible for, the number of individuals with responsibilities increases and the level of responsibility of other individuals rises. This makes it more difficult to assign individual responsibilities and to form collective entities that project unitary instructions and regulations to all individuals involved in the programming, marketing and utilization process.

In this sense, there is no responsibility gap created by the specific type of agency of new AI systems. What emerge are new objects of responsibility, in other words, new things for which we can be responsible. If the scenarios were already difficult with other technologies, where we found numerous individuals and interconnected elements that made the assignment of responsibility challenging, the situation becomes more complicated when we introduce new objects of responsibility. Another interesting issue is the complexity of each of the responsibility objects. As Coeckelbergh [9] points out, one of the fundamental aspects of the problem of many things lies in the different degrees of agency of each of the parts of the technology.

Regarding the second aspect, what Hagendorff [24] has called the AI race is a key consideration: economic and political competition takes AI as a very important asset. The development of these technologies is driven by political and economic interests and a strong technosolutionist narrative [10]. This leads to more and more investments in AI and to it playing an increasingly relevant role in various human activities. As a result, AI has become an infrastructure [27], i.e., a technology that permeates all or almost everything that humans do. AI would not have to be involved in all human activities nor would so much money have to be invested in these technologies, but for societal reasons it is done.

In this sense, the gap produced by new AI systems aggravates the problems of responsibility for two main social reasons: the number of AI systems; and their ubiquity and extension. As we have said before, one of the main excuses for exempting responsibilities refers to circumstances. If we consider the effects of something to be minor or the scope in which those effects are embedded to be of little importance, we may be willing to waive responsibility [49, 68]. The social configuration of new AI systems makes these excuses difficult to admit. On the one hand, numerous AI systems are proliferating, both various types of models and models that have a very significant extent. If the scope and number of such systems were minimal, then there would be no very serious responsibility problems. Their effects would be minimal. This, however, is not what happens. On the other hand, we know that these systems are ubiquitous and are beginning to occupy all human activities [78]. In this sense, they are beginning to act in morally sensitive areas. The medical, educational or judicial spheres are governed by more comprehensive moral requirements and make higher demands for responsibility [38]. In short, if their effects were small and/or they would act in areas of low importance, the responsibility gap would not be very significant.

Another important aspect of the social dimension is the type of relationship we maintain with technology [2]. If we conceive it as something external to us, beyond our reach, we will attribute to it an inevitability that it does not really have. Technology and human beings do not belong to separate spheres but are mutually related and interdependent [76]. In this sense, it may be fruitful to understand our current relationship with technology as a process of alienation [4, 61]: we are able to hold humans responsible for AI, although it may be difficult because of the problem of the many hands, but we do not do so because we believe AI systems to be external and independent entities. As Latour [37] argues, technologies, including AI, are not just finished products, what Latour calls “black boxes”, but sociotechnical objects that include the participation of many human beings and political and economic interests. Political approaches to AI [10] are showing the need to democratize access to the design and values of systems so that they do not remain solely in the hands of engineers and managers of large companies.[21]

4.3 Recapitulation

Where does the argument leave us? Oimann [55] has differentiated three types of discussions within the responsibility gap debate: (i) those that have to do with the existence of

[21] I am indebted to an anonymous reviewer for her/his suggestion to incorporate the concept of alienation into the analysis of the responsibility gap.

the gap, (ii) those that have to do with the novelty of this moral problem, and (iii) those that have to do with ways to address and solve it. In our case, I have argued that the responsibility gap arising from new AI systems does exist (i), but that it is not a new moral problem (ii). It is reducible to the many hands problem and collective agency. However, it has specific internal and external aspects that differentiate it from the emergence of both problems in other fields and disciplines.

With respect to (iii), some considerations are worth making about the possible solutions that derive from this theoretical analysis. First, the increase in the number of objects of responsibility leads us towards some technical solutions. If we are responsible both for the data from which AI is fed and for the ways in which it learns, prevention and traceability tools do need to emerge. Data sheets [22], for example, facilitate better data processing and prevent potential errors and omissions. Traceability, for its part, helps us to know which individuals are causally linked to the different levels and parts of the programming [60]. Second, another type of solution comes from more normative considerations. I have shown how it is possible to take responsibility for other non-full moral agencies. However, we can also take responsibility for the actions of full moral agencies: this is the case of the figure of the respondeat superior [8]. In this sense, the development of AI must be accompanied by clear levels of control and accountability. Third, the problem of ethical proliferation should lead us to rescue proposals and solutions from the economic and political sphere. For example, some authors have explored the possibility of the legal personhood of AI analogous to that of a corporation [7]. Fourth, it is important to take into account the domains in which different AIs are used. As I have shown, there are very sensitive fields that involve greater responsibilities, and this implies that the use of AI in them must be carried out with very low levels of risk and more exhaustive prevention and traceability measures. Finally, the reduction of the responsibility gap to the problem of many hands and collective agency should make us wary of certain social conceptions of technology. They cause us to conceive technology as something separate and alien to us, over which we have no power. Despite the difficulties, I have shown that we have ways of assigning responsibility for the actions and decisions of new AI systems. We must not fall into illusions of technological determinism [2].

5 Conclusion

In this article, I have argued that the responsibility gap arising from new AI systems is reducible to the problem of many hands and collective agency. Systematic analysis of the agential dimension of AI has led me to outline a disjunctive between the two problems. Either we reduce the individual responsibility gaps to the many hands or we abandon the individual dimension and accept the possibility of a responsible collective agency. Moreover, I have argued that this conclusion reveals an underlying weakness in the various discussions of AI ethics: the habitual lack of attention to the question of the disciplinary boundaries of AI ethics. This absence has made it difficult to identify the specifics of the responsibility gap arising from new AI systems as compared to the responsibility gaps of other applied ethics. In the last subsection, I have been concerned with outlining these specific aspects.

The argumentation has proceeded as follows. First, I have developed the general concept of the responsibility gap by attending to both the elements and conditions of responsibility. This has revealed that the responsibility gap arising from the new AI systems depends on the breach of the agential control condition. Subsequently, I have presented the three strategies that have been pursued to address the responsibility gap. While each of them has strengths, I have argued that their weaknesses lead us to conduct a systematic analysis of the agential dimension of new AI systems.

Second, I have analyzed three types of conceptions of AI agency: moral agency with responsibility, moral agency without responsibility, and neither moral agency nor responsibility. The first one leads us to understand AI from the concept of collective agency and the last two are reducible to the many hands problem.

Finally, I have exposed the problem of the disciplinary boundaries of AI ethics and argued that it is at the root of the difficulties that the responsibility gap of new AI systems has generated. After that, I have shown the specific aspects of the responsibility gap produced by these systems: from an internal point of view, their large and complex databases and the criteria used in algorithmic training; and from an external point of view, the increase in the number of systems, their ubiquity and our social conception of technology. This has made it possible to show what is specific to the many hands and collective agency problems in AI ethics.

Acknowledgements I am very grateful to Francisco Lara and Jon Rueda for their insightful and valuable comments. I am also grateful to the attendees of the II Ibero-American Seminar on Technology Ethics, held at the University of Granada; the Novatores Workshop on Philosophy of Artificial Intelligence, held at the University of Salamanca; the Forum on Philosophy, Engineering, and Technology (FPET) 2024, held at the Karlsruhe Institute of Technology; the ESDiT / 4TU.Ethics Conference, held at the University of Twente; and the AAAI/ACM Conference On Artificial Intelligence, Ethics, And Society, held at the San José Convention Center (CA/USA) for their insightful comments.

Author contributions JLA is the sole author of this paper.

Funding Funding for open access publishing: Universidad de Granada/CBUA. The author has benefited from a research contract FPU (FPU22/03178), funded by the Spanish Ministry of Education. This article is also part of the research project AutAI, funded by the State Research Agency of the Spanish Government (PID2022-137953OB-I00).

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The author declares that he has no conflict of interest.

Clinical trial number Not applicable.

References

1. Llorca Albareda, J., Rueda, J.: Divide and rule? why ethical proliferation is not so Wrong for technology ethics. Philos. Technol. **36**, 10 (2023). https://doi.org/10.1007/s13347-023-00609-8
2. Llorca Albareda, J.: Anthropological crisis or crisis in moral status: a philosophy of technology approach to the moral consideration of artificial intelligence. Philos. Technol. **37**, 12 (2024). https://doi.org/10.1007/s13347-023-00682-z
3. Llorca Albareda, J., García, P., Lara, F.: The moral status of AI entities. In: Lara, F., Deckers, J. (Eds.) Ethics of artificial intelligence, The International Library of Ethics, Law and Technology, **41**, 59-83, Springer, Cham (2024). https://doi.org/10.1007/978-3-031-48135-2_4
4. Biondi, Z.: The specter of automation. Philosophia. **51**(3), 1093–1110 (2023). https://doi.org/10.1007/s11406-022-00604-x
5. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford (2014)
6. de Bruin, B., Floridi, L.: The ethics of cloud computing. Sci Eng. Ethics. **23**(1), 21–39 (2017). https://doi.org/10.1007/s11948-016-9759-0
7. Chesterman, S.: Artificial intelligence and the limits of legal personality. Int. Comp. Law Q. **69**(4), 819–844 (2020). https://doi.org/10.1017/S0020589320000366
8. Chomanski, B.: Liability for Robots: Sidestepping the gaps. Philos. Technol. **34**, 1013–1032 (2021). https://doi.org/10.1007/s13347-021-00448-5
9. Coeckelbergh, M.: Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng. Ethics. **26**(4), 2051–2068 (2020). https://doi.org/10.1007/s11948-019-00146-8
10. Coeckelbergh, M.: The Political Philosophy of AI: An Introduction. Wiley, Hoboken (2022)
11. Coeckelbergh, M.: Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making. AI Soc. **38**(6), 2437–2450 (2023). https://doi.org/10.1007/s00146-021-01375-x
12. Constantinescu, M., Voinea, C., Uszkai, R., Vică, C.: Understanding responsibility in responsible AI. Dianoetic virtues and the hard problem of context. Ethics Inf. Technol. **23**, 803–814 (2021). https://doi.org/10.1007/s10676-021-09616-9
13. Courtenage, S.: Intelligent machines, collectives, and moral responsibility. AI Ethics. 1–14 (2023). https://doi.org/10.1007/s43681-023-00285-6
14. Danaher, J.: Robots, law and the retribution gap. Ethics Inf. Technol. **18**(4), 299–309 (2016). https://doi.org/10.1007/s10676-016-9403-3
15. Danaher, J., Nyholm, S.: Automation, work and the achievement gap. AI Ethics. **1**(3), 227–237 (2021). https://doi.org/10.1007/s43681-020-00028-x
16. De Jong, R.: The retribution-gap and responsibility-loci related to robots and automated technologies: a reply to Nyholm. Sci Eng. Ethics. **26**(2), 727–735 (2020). https://doi.org/10.1007/s11948-019-00120-4
17. Demirtas, H.: AI responsibility gap: not new, inevitable, unproblematic. Ethics Inf. Technol. **27**(1), 7 (2025). https://doi.org/10.1007/s10676-024-09814-1
18. Fischer, J.M., Ravizza, M.: Responsibility and Control: A Theory of Moral Responsibility. Cambridge University Press, Cambridge (1998)
19. Floridi, L.: Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. Philosophical Trans. Royal Soc. A: Math. Phys. Eng. Sci. **374**(2083), 20160112 (2016). https://doi.org/10.1098/rsta.2016.0112
20. Floridi, L., Sanders, J.: On the morality of artificial agents. Mind. Mach. **14**, 349–379 (2004). https://doi.org/10.1023/B:MIND.0000035461.63578.9d
21. Frank, L., Klincewicz, M.: Uses and abuses of AI Ethics. In: Gunkel, D. (ed.) Handbook of the Ethics of AI, pp. 205–217. Edward Elgar Publishing, Northampton (2024). https://doi.org/10.4337/9781803926728.00019
22. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumeé, H. III, Crawford, K.: Datasheets for datasets. arXiv: 1–17. (2018). https://doi.org/10.48550/arXiv.1803.09010
23. Gunkel, D.: Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond. MIT Press, Cambridge (2023)
24. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
25. Hakli, R., Mäkelä, P.: Moral responsibility of robots and hybrid agents. Monist. **102**(2), 259–275 (2019). https://doi.org/10.1093/monist/onz009
26. Hanson, F.A.: Beyond the skin bag: on the moral responsibility of extended agencies. Ethics Inf. Technol. **11**(1), 91–99 (2009). https://doi.org/10.1007/s10676-009-9184-z
27. Heilinger, J.C.: The ethics of AI ethics. A constructive critique. Philos. Technol. **35**(3), 61 (2022). https://doi.org/10.1007/s13347-022-00557-9
28. Himma, K.E.: Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? Ethics Inf. Technol. **11**, 19–29 (2009). https://doi.org/10.1007/s10676-008-9167-5
29. Hindriks, F., Veluwenkamp, H.: The risks of autonomous machines: From responsibility gaps to control gaps. Synthese. **201**(1), 21 (2023). https://doi.org/10.1007/s11229-022-04001-5
30. Johnson, D.G.: Technology with no human responsibility? J. Bus. Ethics. **127**(4), 707–715 (2015). https://doi.org/10.1007/s10551-014-2180-1

31. Kiener, M.: AI and responsibility: no gap, but abundance. J. Appl. Philos. (2024). https://doi.org/10.1111/japp.12765
32. Kitchin, R.: Thinking critically about and researching algorithms. Inform. Commun. Soc. **20**(1), 14–29 (2017). https://doi.org/10.1080/1369118X.2016.1154087
33. Köhler, S.: Instrumental robots. Sci Eng. Ethics. **26**(6), 3121–3141 (2020). https://doi.org/10.1007/s11948-020-00259-5
34. Köhler, S., Roughley, N., Sauer, H.: Technologically blurred accountability? Technology, responsibility gaps and the robustness of our everyday conceptual scheme. In: Ulbert, C., Finkenbusch, P., Sondermann, E., Debiel, T. (eds.) Moral Agency and the Politics of Responsibility, pp. 51–68. Routledge, Oxfordshire (2017)
35. Königs, P.: Artificial intelligence and responsibility gaps: What is the problem? Ethics Inf. Technol. **24**(3), 36 (2022). https://doi.org/10.1007/s10676-022-09643-0
36. Lara, F., Deckers, J.: Artificial intelligence as a socratic assistant for moral enhancement. Neuroethics. **13**(3), 275–287 (2020). https://doi.org/10.1007/s12152-019-09401-y
37. Latour, B.: Pandora's hope: essays on the Reality of Science Studies. Harvard University Press (1999)
38. Lauwaert, L.: Artificial intelligence and responsibility. AI Soc. **36**, 1001–1009 (2021). https://doi.org/10.1007/s00146-020-01119-3
39. List, C.: Group agency and artificial intelligence. Philos. Technol. **34**(4), 1213–1242 (2021). https://doi.org/10.1007/s13347-021-00454-7
40. List, C., Pettit, P.: Group Agency: The Possibility, Design, and Status of Corporate Agents. Oxford University Press, Oxford (2011)
41. Loh, J.: Responsibility and robot ethics: a critical overview. Philosophies. **4**(4), 58 (2019). https://doi.org/10.3390/philosophies4040058
42. Loh, W., Loh, J.: Autonomy and responsibility in hybrid systems. Robot ethics, 2. (2017)
43. Martin, K.: Ethical implications and accountability of algorithms. J. Bus. Ethics. **160**, 835–850 (2019). https://doi.org/10.1007/s10551-018-3921-3
44. Matthias, A.: The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf. Technol. **6**(3), 175–183 (2004). https://doi.org/10.1007/s10676-004-3422-1
45. McKenna, M.: Conversation and Responsibility. Oxford University Press, Oxford (2012)
46. Mirzaeighazi, S., Stenseke, J.: Responsibility before Freedom: Closing the responsibility gaps for autonomous machines. AI Ethics. 1–13 (2024). https://doi.org/10.1007/s43681-024-00503-9
47. Moor, J.H.: The nature, importance, and difficulty of machine ethics. IEEE. Intell. Syst. **21**(4), 18–21 (2006). https://doi.org/10.1109/MIS.2006.80
48. Munn, L.: The uselessness of AI ethics. AI Ethics. **3**(3), 869–877 (2023). https://doi.org/10.1007/s43681-022-00209-w
49. Neuhäuser, C.: Some sceptical remarks regarding robot responsibility and a way forward. In: Misselhorn, C. (ed.) Collective Agency and Cooperation in Natural and Artificial Systems, pp. 131–146. Springer, New York (2015). https://doi.org/10.1007/978-3-319-15515-9_7
50. Nickel, P.J., Franssen, M., Kroes, P.: Can we make sense of the notion of trustworthy technology? Knowl. Technol. Policy. **23**, 429–444 (2010). https://doi.org/10.1007/s12130-010-9124-6
51. Nissenbaum, H.: Accountability in a computerized society. Sci Eng. Ethics. **2**, 25–42 (1996). https://doi.org/10.1007/BF02639315
52. Nyholm, S.: Attributing agency to automated systems: reflections on human–robot collaborations and responsibility-loci. Sci Eng. Ethics. **24**(4), 1201–1219 (2018). https://doi.org/10.1007/s11948-017-9943-x
53. Nyholm, S.: This is Technology Ethics: An Introduction. Wiley, Hoboken (2022)
54. Nyholm, S.: Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. In: Placani, A., Broadhead, S. (eds.) Risk and Responsibility in Context, pp. 191–213. Routledge, Oxfordshire (2023). https://doi.org/10.4324/9781003276029-14
55. Oimann, A.K.: The responsibility gap and LAWS: A critical mapping of the debate. Philos. Technol. **36**, 3 (2023). https://doi.org/10.1007/s13347-022-00602-7
56. Oimann, A.K., Tollon, F.: Responsibility gaps and technology: Old wine in new bottles? J. Appl. Philos. (2024). https://doi.org/10.1111/japp.12763
57. Pettit, P.: Responsibility incorporated. Ethics. **117**(2), 171–201 (2007). https://doi.org/10.1086/510695
58. Rudy-Hiller, F.: The epistemic condition for moral responsibility. Stanf. Encyclopedia Philos., (2018). https://plato.stanford.edu/entries/moral-responsibility-epistemic/
59. Sætra, H.S., Danaher, J.: To each technology its own ethics: the problem of ethical proliferation. Philos. Technol. **35**(4), 1–26 (2022). https://doi.org/10.1007/s13347-022-00591-7
60. Santoni de Sio, F., Mecacci, G.: Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philos. Technol. **34**, 1057–1084 (2021). https://doi.org/10.1007/s13347-021-00450-x
61. Sidorkin, A.M.: Embracing liberatory alienation: AI will end us, but not in the way you may think. AI Soc. 1–8 (2024). https://doi.org/10.1007/s00146-024-02019-6
62. Simpson, T.W., Müller, V.C.: Just war and robots' killings. Philosophical Q. **66**(263), 302–322 (2016). https://doi.org/10.1093/pq/pqv075
63. Sison, A.J.G., Redín, D.M.: A neo-aristotelian perspective on the need for artificial moral agents (AMAs). AI Soc. **38**(1), 47–65 (2023). https://doi.org/10.1007/s00146-021-01283-0
64. Shoemaker, D.: Responsibility from the Margins. Oxford University Press, Oxford (2015)
65. Sparrow, R.: Killer robots. J. Appl. Philos. **24**(1), 62–77 (2007). https://doi.org/10.1111/j.1468-5930.2007.00346.x
66. Strawson, P.F.: Freedom and resentment and other essays. Proceedings of the British Academy 48:1–25. (1962)
67. Thompson, D.F.: Moral responsibility of public officials: the problem of many hands. Am. Polit. Sci. Rev. **74**(4), 905–916 (1980). https://doi.org/10.2307/1954312
68. Tigard, D.W.: There is no techno-responsibility gap. Philos. Technol. **34**(3), 589–607 (2020). https://doi.org/10.1007/s13347-020-00414-7
69. Tigard, D.W.: Artificial moral responsibility: How we can and cannot hold machines responsible. Camb. Q. Healthc. Ethics. **30**(3), 435–447 (2021a). https://doi.org/10.1017/S0963180120000985
70. Tigard, D.W.: Responsible AI and moral responsibility: a common appreciation. AI Ethics. **1**(2), 113–117 (2021b). https://doi.org/10.1007/s43681-020-00009-0
71. Tigard, D.W.: Workplace automation without achievement gaps: a reply to Danaher and Nyholm. AI Ethics. **1**(4), 611–617 (2021c). https://doi.org/10.1007/s43681-021-00064-1
72. Tollon, F.: Responsibility gaps and the reactive attitudes. AI Ethics. **3**, 295–302 (2023). https://doi.org/10.1007/s43681-022-00172-6
73. Totschnig, W.: The problem of superintelligence: political, not technological. AI Soc. **34**(4), 907–920 (2019). https://doi.org/10.1007/s00146-017-0753-0
74. Van de Poel, I.: The problem of many hands. In: Van de Poel, I., Royakkers, L., Zwart, S.D. (eds.) Moral Responsibility and the

Problem of many Hands, pp. 50–92. Routledge, Oxfordshire, pp 50–92 (2015)

75. Vallor, S., Vierkant, T.: Find the gap: AI, responsible agency and vulnerability. Mind. Mach. **34**(3), 20 (2024). https://doi.org/10.1007/s11023-024-09674-0
76. Verbeek, P.P.: What Things do: Philosophical Reflections on Technology, Agency, and Design. Penn State (2005)
77. Watson, G.: Two faces of responsibility. Philosophical Top. **24**(2), 227–248 (1996). https://doi.org/10.5840/philtopics199624222
78. Weiser, M.: The computer for the 21st Century. Sci. Am. **3**(265), 94–104 (1991). https://doi.org/10.1145/329124.329126
79. Williams, G.: Responsibility as a virtue. Ethical Theory Moral. Pract. **11**, 455–470 (2008). https://doi.org/10.1007/s10677-008-9109-7
80. Winner, L.: The Whale and the Reactor: A Search for Limits in an age of high Technology. University of Chicago Press, Chicago (1986/2010)
81. Zimmerman, M.J.: Moral responsibility and ignorance. Ethics. **107**(3), 410–426 (1997). https://doi.org/10.1086/233742

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.