Minds and Machines (2025) 35:12
https://doi.org/10.1007/s11023-025-09712-5

The Testimony Gap: Machines and Reasons

Robert Sparrow[1] · **Gene Flenady**[1]

Received: 13 February 2024 / Accepted: 19 January 2025

Abstract
Most people who have considered the matter have concluded that machines cannot be moral agents. Responsibility for acting on the outputs of machines must always rest with a human being. A key problem for the ethical use of AI, then, is to ensure that it does not block the attribution of responsibility to humans or lead to individuals being unfairly held responsible for things over which they had no control. This is the "responsibility gap". In this paper, we argue that the claim that machines cannot be held responsible for their actions has unacknowledged implications for the conditions under which the outputs of AI can serve as reasons for belief. Following Robert Brandom, we argue that, because the assertion of a claim is an action, moral agency is a necessary condition for the giving and evaluating of reasons in discourse. Thus, the same considerations that suggest that machines cannot be held responsible for their actions suggest that they cannot be held to account for the epistemic value — or lack of value — of their outputs. If there is a responsibility gap, there is also a "testimony gap." An under-recognised problem with the use of AI, then, is to ensure that it does not block the attribution of testimony to human beings or lead to individuals being held responsible for claims that they have not asserted. More generally, the "assertions" of machines are only capable of serving as justifications for belief or action where one or more people accept responsibility for them.

Keywords Artificial intelligence · Ethics · Testimony · Brandom · Wittgenstein · Responsibility

✉ Robert Sparrow
robert.sparrow@monash.edu

[1] Department of Philosophy, School of Philosophical, Historical and International Studies, Faculty of Arts, Monash University, Clayton, VIC, Australia

Published online: 24 February 2025

We live in an age of intelligent machines. Recent progress in the development and applications of artificial intelligence means that many citizens of highly industrialised nations now deal with technologies that contain at least some element of AI on a daily basis. Increasingly, important medical, legal, administrative, and even political decisions are being made by AI. Indeed, as driverless vehicles appear on our roads, and autonomous weapon systems are deployed in war, it is no exaggeration to say that some of the decisions that machines are making are matters of life and death. Even when human lives are not at risk, the outputs of AI may be high stakes, such as when, for instance, they decide who gets confined to prison or who is denied a housing loan.

It is little wonder, then, that the question as to who should be held responsible for the "decisions" of AI has quickly moved to the forefront of ethical and regulatory debate about AI (Pasquale, 2015; High-Level Expert Group on Artificial Intelligence, 2019; Jobin et al., 2019). Those who have been harmed by machines may wish to seek compensation. Those who have been wronged by machines may want to know who is to blame. However, most critics who have considered the matter have concluded that the responsibility for the actions of machines must always rest with a human being (Johnson, 2006; Véliz, 2021). Machines cannot be full moral agents. That is to say, machines can never be held morally responsible for the consequences of their outputs. A key problem for the ethical use of AI, then, is to ensure that the use of AI does not block the attribution of responsibility to human beings or lead to individuals being unfairly held responsible for things over which they had no control. This is the notorious "responsibility gap" (Matthias, 2004; Sparrow, 2007).

In this paper we argue that the claim that machines cannot be held responsible for their actions has unacknowledged implications for the conditions under which the outputs of AI can serve as reasons for belief. If there is a responsibility gap, then there is also a "testimony gap". In Sect. 1, "Machines, agency, and embodiment", we note the near consensus that machines are not moral agents and cannot be held responsible for the consequences of their outputs and we rehearse the arguments that we think best justify these conclusions. In Sect. 2, "Brandom on the pragmatics of judgment" we summarise Robert Brandom's argument, first articulated in *Making it Explicit* (1994) but set out more recently in *Articulating Reasons* (2000) and *Reason in Philosophy* (2009), that, because the assertion of a claim is an action, moral agency is a necessary condition for the giving and evaluating of reasons in discourse. The quality of one's testimony is something for which one can be held responsible. Indeed, it is a condition of giving testimony at all that we "stand behind our words." Thus, as we discuss in Sect. 3, "The testimony gap", the same considerations that suggest that machines cannot be held responsible for their actions in general suggest that they cannot be held to account for the epistemic value — or lack of value — of their claims. An under-recognised problem with the use of AI, then, is to ensure that it does not block the attribution of testimony to human beings or lead to individuals being held responsible for claims that they have not asserted. In Sect. 4, "Machines and reasons", we provide a diagnosis as to why the existence of a testimony gap has not been more widely recognised and argue that the "assertions" of "intelligent" machines are only capable of serving as justifications for belief or action where one or more people accept responsibility for them. Section 5, "Living with the Testimony Gap" discusses some ways in which the burden of accepting this responsibility may

be relieved and the testimony gap plastered over, if not filled. We conclude by suggesting that, short of abandoning technological society, we are doomed to live with the testimony gap. Acknowledging this fact, though, has the potential to open up new possibilities for the form such a society takes.

1 Machines, Agency, and Embodiment

As we noted above, it is an article of faith in most discussions of AI ethics that machines cannot be moral agents: responsibility for the consequences of the "decisions" of AI always comes back to one or more human beings. Here, for instance, are some representative discussions of the matter, which, although they disagree on the reasons why, concur that machines cannot be moral agents:

> Despite the lack of agreement concerning what the exact conditions of moral agency are, the majority view is that present-day robots and AI agents do not satisfy them. An agent must satisfy the conditions of moral agency in order to qualify as a proper bearer of moral responsibility. Hence, according to this line of thinking, current robots and AI agents are not fit to be held morally responsible.
>
> Hakli and Mäkelä (2019), 260.

> … computer systems do not and cannot meet one of the key requirements of the traditional account of moral agency. Computer systems do not have mental states and even if states of computers could be construed as mental states, computer systems do not have intendings to act arising from their freedom. Thus, computer systems are not and can never be (autonomous, independent) moral agents.
>
> Johnson (2006), 203–204.

> When algorithms cause moral havoc, as they often do, we must look to the human beings who designed, programmed, commissioned, implemented, and were supposed to supervise them to assign the appropriate blame. For all their complexity and flair, algorithms are nothing but tools, and moral agents are fully responsible for the tools they create and use.
>
> Véliz (2021), 493.

> Legal and moral responsibility for a robot's actions should be no different than they are for any other AI system, and these are the same as for any other tool. Ordinarily, damage caused by a tool is the fault of an operator, and benefit from it is to the operator's credit. If the system malfunctions due to poor manufacturing, then the fault may lay with the company that built it, and the operator can sue to resolve this… We should never be talking about machines taking ethical

> decisions, but rather machines operated correctly within the limits we set for them."
>
> Bryson (2010), 69.

Such examples could be multiplied with ease.

Our own diagnosis of the reasons why machines cannot be held morally responsible, which one of the authors has developed at length elsewhere, points to the relationship between key concepts surrounding the idea of responsibility and our affective responses to the embodied presence of other agents — and to the nature of the "bodies" of machines (Sparrow, 2004; Sparrow, 2007; Sparrow, 2021. See also: Cockburn, 1990; Gaita, 2004; and Gleeson, 2001). In order for someone or something to be capable of being held morally responsible, it must be conceivable that the entity could feel remorse for its actions and be an appropriate object of punishment: it must also be the case that *we* could imagine feeling remorse for having held the entity responsible for a putative wrong unjustly. These are conceptual ("grammatical") claims — claims about how these concepts "work" — rather than claims about what must be true empirically at any particular moment in a particular case (Wittgenstein, 1989, § 90).

The difficulty with attributing responsibility to machines arises because the conditions of application of these concepts (remorse, punishment, desert) in turn make essential reference to our own affective states and, by implication, to the affective — and, importantly, bodily — states of machines. A key insight, developed in the course of Wittgenstein's analysis of the problem of other minds in the *Philosophical Investigations*, is that our ability to make claims about the thoughts and feelings of others relies, to a large degree, on emotions and affective responses that are expressed on — and through — the human face and body (Cockburn, 1985, 1990; Gaita, 2004, 164–188; Winch, 1980; Wittgenstein, 1989, § 243–427, II (iv). For example, in order to be able to determine whether what someone who has committed a moral wrong is experiencing and expressing is remorse, rather than, for instance, irritation with the consequences of their actions, we must make reference to their facial expressions and tone of voice when they speak about the matter. Where it is impossible in principle to make these sorts of subtle judgments about affect, the distinction between remorse and irritation has no content. More generally, our capacity to attribute emotions and other mental states to other human beings — and, perhaps, some animals — relies upon our ability to make a distinction between genuine instances and false semblances of such, which in turn is ultimately grounded in our ability to "see" feelings in the face of other people, or in the movements of their bodies (Cockburn, 1994; Gaita, 1999, 237–258; Gleeson, 2001; Winch, 1980; Wittgenstein, 1989, § 281–287).

According to this line of thinking, then, the problem with attributing responsibility to machines is that they lack expressive faces and bodies of the sort that would allow us to tell whether we are using the language surrounding responsibility appropriately, and, in particular, to ground a distinction between reality and appearance when it comes to the feelings of machines. What would it mean to say of something that looked like a filing cabinet, that it was "feeling remorse" (Sparrow, 2007, 2021; Gaita, 2004; 279–282)? How would we discern whether the claim was true? If we

wanted to punish a machine, how would we know whether we had been successful? Insofar as the idea of punishment — as opposed to rehabilitation — is linked to the idea of suffering, unless machines are capable of suffering, they will not be appropriate objects of punishment (Asaro, 2012). However, how would we tell whether a machine was suffering? The bodies of machines are inscrutable to us. No amount of "behaviour" by a machine will suffice to establish that it is really feeling, rather than pretending to feel, something. Conversely, if someone claimed that they were feeling remorse for having treated a machine unjustly, how would we evaluate the claim? The sorts of emotional responses and relationships that would ordinarily give content to the claim are not possible where their "object" lacks the sort of moral individuality that is characteristic of human beings and that grounds the idea that we owe other people justice (Gaita, 2004, 141–163).

The detail of the argument above is properly controversial. In particular, tying the capacity for moral agency to the possession of a certain sort of body flies in the face of what is arguably the historically dominant philosophical tradition of thought about the relationship between bodies and minds, which is essentially Cartesian insofar as it denies that there is any necessary connection between mental states and bodily affect. It would also appear to have important implications for moral philosophy more generally, which are explored in a literature on the implications of Wittgenstein's thought for ethics (Gaita, 2004; Cockburn, 2021; Pianalto, 2011). It is, therefore, important to emphasise that, for the purposes of our larger argument, it does not matter whether the reader finds our own preferred diagnosis of the reasons *why* machines cannot be full moral agents compelling: all that matters is that he/she agrees that machines cannot be full moral agents.

2 Brandom on the Pragmatics of Judgment

That machines cannot be moral agents has implications for the conditions under which the outputs of machines can provide us with reasons for belief and, therefore, for action, which have not to date been widely recognised on the literature on the epistemology, or the ethics, of AI. In order to draw out these implications, it will prove useful to provide a brief summary of the account of rational agency developed by the pragmatist philosopher Robert Brandom.

Brandom's philosophical project, introduced in *Making it Explicit* (1994) and continually developed in subsequent works (Brandom, 2000, 2002, 2009, 2019), is to advance an "inferentialist" semantic theory against what he terms the "representationalist" account of conceptual content. Brandom takes representationalism to be the dominant paradigm in modern Western philosophy; it understands the meaningfulness of propositions as derived from mental states that represent or "picture" objects and events in the world (Brandom, 1994, p. 73; 2002, 30; see also Redding, 2007, 56–69). In contrast, Brandom's inferentialism self-consciously develops Wilfrid Sellars argument in *Empiricism and the Philosophy of Mind* that a knowledge claim is not an "empirical description" of some mental "episode or state," but rather the placing of that claim in a "logical space of reasons," that is, in inferential relations to other possible knowledge claims (Sellars, 1956/1997, § 36). For Sellars, as for Brandom,

knowledge is essentially a question of *justifying* one's claim to know by properly tracking its inferential implications (Sellars, 1956/1997, § 36). Brandom elaborates on and systematises Sellars' position by arguing that the meaning of a judgment is determined by its standing in inferential relations of "material incompatibility" and "material consequence" with other possible judgments (Brandom, 2019, p. 3). For Brandom, these relations of incompatibility and consequence come in two kinds, corresponding to the "objective" and "subjective" sides of the activity of judgment. On the side of "objective states of affairs," it is *impossible* for materially incompatible states of affairs to obtain, or for a state of affairs to obtain without its material consequent. In contrast, on the side of "subjective thought," it is *not impossible* but, rather, *wrong* for agents to make incompatible claims or to fail to draw a claim's necessary consequents (2009, 48; 2019, 3–4). As such, Brandom's semantic theory is constitutively *normative*: agents *ought not* endorse incompatible claims or endorse some claim while failing to acknowledge its consequents (2009, 48; 2019, 4). For example, if one claims that an object is wholly made of copper, one ought to further endorse the consequent claim that it is an electrical conductor, just as one ought to deny that the object is wholly made of aluminium (Brandom, 2019, p. 2). One grasps the meaning of "copper" not by putatively describing some inner representation, but by properly grasping what other claims the judgment of something being copper *ought* to entail and exclude.

The constitutive normativity of this account of conceptual content obliges Brandom to provide an account of the intersubjective or the social: as he puts it, "because the space of reasons is a *normative* space, it is a *social* space" (2009, 4). For Brandom, both the judgments and the intentional actions of "sapient" creatures (human animals) are to be distinguished from the stimuli responses of nondiscursive, merely "sentient", creatures (nonhuman animals) insofar as judgments are, just as are other actions, loci of *responsibility* (2009, 32, 2019, 9). When human agents judge or act, they are implicitly taking responsibility for justifying that judgment or action to other agents.

Brandom routinely characterizes both judgments and actions as *commitments* (1994, 2000, 2002, 2009, 2019), in the sense that judgments, as are actions more generally, are "performances" by which the agent commits themselves to a claim about "how things are" (in a judgment) or "how things will be" (in an action), and, implicitly, commits to defending that claim with reasons (2019, 9). In other words, in judging, as in acting, one is *taking* the world and oneself to be a certain way, and one is thus implicitly responsible for justifying that "take" to other reason-responsive agents (2002, 21; 2009, 32). Thus, our "normative status" (our "commitments and entitlements," or what we are responsible for and have "authority" to say and do) are for Brandom dependent the "normative attitudes" of *agents* who "*hold* each other responsible, *acknowledge* each other's responsibility, [and] *attribute* commitments and entitlements" (2009, 4, 61; 2019, 298–300).

Brandom has termed this social activity of attributing normative statuses and holding agents to them "deontic scorekeeping" (1994, 2000), but more recently has developed this picture in terms of G.W.F. Hegel's account of reciprocal "recognition" [*Anerkennung*] (2002, 2009, 2019). To take a helpful example from *Reason in Philosophy*, I might "adopt a certain attitude toward myself," taking myself to be

a good chess player, but whether I have the *status* of "good chess player" – and the accompanying responsibilities and authority within relevant communities – depends on my being recognised as such by others whom I myself recognise as having the authority to decide that status, that is, those others I consider to be good at chess (2009, 70–71). Should the attribution of a particular normative status be withheld by those I recognise as entitled to confer it, I ought to revise or abandon the attitude in question (perhaps I am *not* good at chess, after all).

Brandom takes the social dimension of his inferentialism to be broadly in line with the pragmatist dictum that "use determines meaning" (2009, 12; 2019, 3), insofar as the meanings of our judgments and other actions – and the normative statuses they confer – are dependent on the social practice of "giving and asking for reasons" (2009, 119). Particularly important for our purposes in this paper is the way in which Brandom's inferentialism undoes what he reveals to be arbitrary distinctions between the "cognitive" and the "practical" (Brandom, 2019, p. 11). As we have just shown, Brandom understands judgments, as he does other acts, as modes of constitutively normative – and thus irreducibly social – commitment. As he puts it, sapient creatures are "subject to normative assessment of the extent to which what they think and do accords with their commitments or responsibilities" (2009, 11). In other words, *asserting* that such-and-such is the case is always to *do* something.

The upshot of Brandom's bringing together of judgments and actions as species of commitment might be illustrated as follows. Our actions, as well as our *willingness* to act in various ways, play a central role in the evaluation of our assertions. As both assertions and actions are commitments, we ought to ensure that our assertions *and* our actions are materially compatible. That is, we grasp the meaning of a judgments only insofar as we understand what other judgments *and actions* that judgment would entail and exclude. Brandom offers the helpful example of a two-year old who confidently enters a room and claims "The house is on fire" but who does not grasp the content of their utterance – cannot really mean what they say – because they do not grasp the actions entailed by that utterance; that is, they are not about to grab their favourite toy and flee (Brandom, 2002, p. 360). If we are not willing to back up our words with inferentially entailed actions then we may be judged not to mean them, which in turn implies that others need not take them seriously. That is to say, the assessment of assertions has an implicitly or explicitly *normative* dimension, and such assessment takes in a rational agent's further judgments as well as their other actions. We can be lauded or criticised for our knowledge claims, feel shame at having missed an obvious detail, be informally excluded from future discussions on a particular matter, or, as in cases of professional malpractice, be formally punished for egregious epistemic error. Our acts of judgment are morally evaluable just as our acts more generally are morally evaluable.

3 The Testimony Gap

If Brandom is correct that judgments are acts for which agents are responsible and — as we have argued here and is widely believed — machines cannot be held morally responsible for their outputs, then it follows that machines cannot assert claims.

A machine's "statement" that such-and-such is the case is not something for which it can be held responsible and thus cannot count as testimony.[1] Correspondingly, in-and-of-themselves, "assertions" made by machines provide us with no reason for belief or action.

Our own account of the reasons why machines are not moral agents permits a particularly nice expression of the fundamental problem. Because they lack bodies of the sort that allow claims about punishment, remorse, and moral individuality to have sense, machines cannot "stand behind their words" (Gaita, 1989, 136–140; Gaita, 2004, 268–273, 279; Sparrow, 2021; Taylor, 2014). However, again, it is worth emphasising that, if Brandom is correct, it will follow from any account that denies that machines can be responsible for their actions that they cannot testify, give evidence, or even state claims.

The belief that machines cannot be held responsible for what they do has generated discussion in the literature as to whether this leads to, or risks leading to, a morally problematic "responsibility gap" when it comes to the consequences of the use of AI (Matthias, 2004; Sparrow, 2007). If machines themselves cannot be held responsible, then there are only two other options: either responsibility rests with one or more human beings or there is no one who is responsible. If the latter is true, then the use of AI would be highly ethically problematic. However, holding a human being responsible for consequences of the outputs of an AI also seems problematic in circumstances where they have little control over, or even ability to anticipate, the outputs of the machine.

Our discussion here suggests that this "responsibility" gap generates – or, perhaps, is itself simultaneously – a "testimony gap". If "statements" by machines cannot provide us with reasons for belief or action, then there are only two options. Either a human being must vouch for the machine – must testify on its behalf – or its outputs are evidentially worthless. If the latter is true, then clearly its use would be highly problematic. However, holding a human being responsible for the "testimony" of an AI also seems highly problematic in circumstances where they have little ability to verify the outputs of the machine. This may be the case, for instance, if the AI relies on machine learning such that the (faux) sentences it generates are highly sensitive to the details of the data it encounters in the course of its training and/or operations (Matthias, 2004). It is also likely to be the case more generally given the limits of our own epistemic powers and the sorts of questions that we are calling on AI to answer for us.

As even this brief discussion suggests, questions concerning responsibility and AI have been the topic of sustained philosophical discussion. However, surprisingly little attention has been paid to the consequence of the consensus in this discussion — that machines cannot be held responsible — for the epistemic status of the "assertions" of machines.[2] This is despite the fact that intuitions that point to the existence of a testimony gap are widespread in cases where the outputs of AI are proposed to

[1] For a subtle and informative discussion of the role played by responsibility in testimony, see Moran (2006).

[2] An important exception here, which covers some of the same terrain as the current paper, is Heinrichs and Knell (2021).

be used to make important, or even life or death, decisions. For instance, in the debate about the ethics of the development and deployment of autonomous weapon systems, there has been a significant backlash against the idea that machines might be granted the authority to "decide" who to kill (Ipsos, 2021; Open Roboethics Initiative, 2015; Stop Killer Robots Coalition, 2024). Similarly, in the literature on the medical uses of AI, people are typically very anxious to insist that responsibility for diagnosis or treatment decisions informed by AI should remain with a human being (Smith, 2021; Rajpurkar et al., 2022; Wang et al., 2023).

When we think about these higher-stakes cases, we have a tendency to adopt language that emphasises the "agency" of the AI at the expense of noticing that what the AI does is often just to "tell" human beings to do something by "asserting" that some state of affairs exists. For instance, in the debate about "killer robots", the drama of the act of killing distracts from the fact that it is handing over the *decision* to kill that is morally problematic (Umbrello, 2022, p. 125). As several participants in this debate have pointed out, it would not render the involvement of AI in war any less problematic were they used solely to provide lists of targets to be attacked to people who then attacked them as a matter of routine (Renic & Schwarz, 2023; Schwarz, 2021; Sparrow & Henschke, 2023). What looks like a responsibility gap in one light is clearly a testimony gap in another.

The "testimony gap" has been sorely neglected. In the next section we proffer an explanation for this neglect, which in turn will point towards an account of the implications of the responsibility gap for the relationship between people and AI, which we develop in Sect. 5.

4 Machines and Reasons

The main reason, we believe, that the existence of the testimony gap has gone unrecognised is that in almost every circumstance in which we encounter, and (apparently) rely on machines, another human being has already "vouched for" the machine. That is, a human being must assume responsibility – or be held responsible – *for* the epistemic quality of the machine's output. We think that we are relying on the machine, when we are really relying on another human being. The testimony gap arises earlier in the chain of justification that allows us to act as though machines are a secure foundation for belief and so we do not notice it.

To fully comprehend the extent of the difference in the evidential weight of the testimony of humans and the outputs of machines, and the extent to which, in our relations with machines, we are actually relying on the testimony of other human beings, it is helpful to consider an example that Brandom discusses at several points to explicate the full extent of what is involved in judgment – that of a parrot that has been trained to squawk "Rawk! That is red!" when confronted with a red thing… and does so reliably.[3] Brandom concedes that a parrot, which he insists is not a moral agent, may possess – what he calls – a "reliably differential responsive disposition" (or RDRD) to environmental stimuli (Brandom, 2002, p. 350). However, for Brandom,

[3] See also his discussion of the status of the claims of "idiot savants" in Brandom, 2002, 367.

the difference between the parrot's squawk and a human utterance is that the latter expresses that response as a judgment, and, as such, can be taken up inferentially as a premise from which to draw conclusions (Brandom, 1994, p. 89; Brandom, 2009, p. 118; Brandom, 2019, p. 113). In saying "that is red" the human being applies a concept, and that application commits the individual, as we discussed in Sect. 2 above, to inferential entailments and exclusions. For example, the judgment that such a thing is red commits one further to the claim that it is coloured but excludes one taking that same thing to be green – and so on (Brandom, 1994, p. 89; Brandom, 2009, p. 184; Brandom, 2019, p. 140). In contrast, while the trained parrot "might share reliable differential dispositions with a genuine observer of red things… it is functioning at most as a measuring instrument, labelling, not describing, the things it responds to as red." (Brandom, 2019, p. 113).[4]

In a footnote to his discussion of this and similar cases in *Reason in Philosophy*, Brandom observes that this means that there is a profound difference between what we can conclude on the basis of the parrot's RDRD and on the basis of human "*testimony*" (italics his). Brandom allows that we can infer from the parrot's squawk that there is something red in his field of vision "because the two sorts of events are reliably correlated," just as the activation of a photocell and certain electromagnetic frequencies are reliably correlated (2009, 208 n.9). However, the parrot's squawk does not support the full range of inferences that are justified when a human being reports that the object in front of them is red. When a human being testifies to the presence of redness "What he [another sapient creature] says is usable as a premise in our own inferences, not just the fact that his saying it is reliably correlated with the situation he (but not the parrot) reports (though they both respond to it)."

The outputs of AI are, we submit, like the squawks of Brandom's parrot. They may correspond with some state of affairs but they neither represent these states of affairs nor describe them: they lack conceptual content altogether.[5] For instance, a human being may observe a machine's outputs over time, conclude that they track some phe-

[4] Brandom's discussion uses the example of colour concepts to emphasise the way in which making claims commits us to other applications of these and related concepts. It is important to emphasise that the connections between our concepts can ramify quickly and widely such that the proper application of any concept may have implications for action as well as belief. Thus, this account would not be controverted by machines that *were* capable of tracing out the connections of concepts in a manner that tracked the usages of human beings so long as the machines' ability to do so trails off when it comes to practical commitments. For instance, OpenAI's ChatGPT is already, at the time of writing, capable of "reporting" (sometimes) that if something is red then it is coloured and that if something is red then it cannot be green at the same time. What it cannot do is stop when a light turns red or be held responsible if it should fail to do so - and this has implications for its ability to possess the concept "red" in the first place. Empirical description is already for Brandom "a move in a game of giving and asking for reasons," which, as we have emphasised here, may require that agents be willing to back up their words with actions.

[5] For an alternative, book length, treatment of the semantic and epistemic status of the outputs of AI, which does credit them with conceptual content, see Cappelen & Dever, 2021. It would take us too far afield from our current purpose, of setting out the inferentialist case against the outputs of machines having conceptual content and exploring its implications for questions of responsibility, to respond to the arguments of their book here.

nomenon, and form and assert beliefs on the basis of its outputs but responsibility for so doing rests with them.[6] If the machine is wrong, it cannot be blamed for being so.

The "vouching" of a human being can transform the *mere* output of an AI into a judgment – and thus something that might serve as a justification – by bringing that output into the "space of reasons" as a commitment they (the human being) might have to answer for. In other words, the activity of assessing a content for correctness, that is, normatively assessing the justification of a particular claim, must be "outsourced" from machine to human. Once one (or more) human being(s) has brought the outputs of machines into the space of reasons in this way – has "vouched for" a machine – others may be justified in relying on the machine's outputs in the justification of their beliefs.

Nevertheless, a testimony gap remains. It seems that the vouching agent's epistemic duties cannot be discharged: if their testimony is disputed, then, even though they can offer in their defence that "the machine said so", the responsibility for the epistemic quality of the outputs remains with the human rather than the machine.[7]

This is the form that the testimony gap takes when someone – for instance a designer, a manufacturer, or, perhaps, a regulator – *wants* to vouch for the reliability of a machine. Of course, the testimony gap may also exist, and be even more problematic, when, because *we* want to rely on machines, we simply *hold* someone responsible for the accuracy of their outputs despite the fact that this person is not willing to vouch for them or, perhaps, even capable of vouching for them. In such cases, as with the more familiar "responsibility" gap, the issue arises about the justice of our doing so. That is, we may come to hold people unjustly responsible for claims that they did not, or even could not, stand behind, for instance because they have no knowledge of how the machine has arrived at a particular output.

5 Living with the Testimony Gap

Unless we wish to confront the prospect of radically revising the epistemic practices that underpin our technological society, we need to find ways of living with the testimony gap. In particular, we need to determine the conditions under which, what we have argued is, *a prima facie* inadequate justification – "the AI said so" – can *count* for other rational agents as legitimate.

We believe that Brandom's social-historical model of normativity offers important resources here. As Brandom argues at length in his recent reconstruction of Hegel's

[6] In practice, given the complexity of machines and their use cases, this will often in turn involve relying on the testimony of other human beings when it comes to, for instance, the performance of the machine's components or an account of the circumstances in which the machine may be expected to function properly.

[7] John McDowell – a longtime colleague of Brandom's at the University of Pittsburgh – would put this point in terms of the distinction between a "justification" and an "exculpation." If, when challenged, the voucher for some machine output simply responds, "the machine said so," they have not *justified* the statement they are attributing to the machine, but sought rather to *exculpate* themselves from their responsibility as voucher. A potentially legitimate line of justification would require giving reasons for their commitment to the reliability of the machine's outputs. See McDowell, 1996, 6–21.

philosophy in *A Spirit of Trust* (2019), and as we have discussed briefly above, the attribution of normative statuses is dependent on social relations of reciprocal recognition. A necessary condition of rational agency – being the kind of creature capable of normatively answerable judgment and action – is that we are recognised as such, and that we reciprocally confer the status of rational agency on those who so recognise us (2019, 260). As we discussed in Sect. 2 above, to claim or to act is to make a commitment to justify that claim or action and, at the same time, to implicitly take others to be entitled to assess those reasons. However, it is only insofar as I recognise your authority to contest my epistemic claims, and open myself up to possible disagreement, even criticism, dismissal, and censure, that your assent to those claims can be meaningful and entitle me to exercise authority through those claims. If I did not take others to be authoritative in this way, my claims would lack rational legitimation and thus, as we have discussed in detail, determinate conceptual content. Grasp of a content requires the capacity to draw the right inferences, both "cognitive" (an entailed further claim) and "practical" (an implicated action): the "right" inferences are established over time by a particular community of mutually recognising agents. Thus, Brandom, like Hegel, takes this structure of reciprocal recognition to ground, not just the status of agents as rational, but the capacity of agents to make judgments at all. We determine collectively and historically what can count as a legitimate justification for our commitments.

Brandom's favourite example – and indeed underlying model – for this account of collective normative self-determination is British and American common or "judge-made" law. At common law, the judge in a particular case determines the norm governing the facts at hand, but in so doing they undertake a commitment for which they are responsible at once to past and future. Should the judge in the present case decide to revise the norm's content and/or its conditions of applicability and so depart from historical precedent, they are under an obligation to justify to future judges that departure – on pain of their decision being retroactively revoked as illegitimate (Brandom, 1994, p. 130; Brandom, 2000, p. 76; Brandom, 2002, 13–14; Brandom, 2009, 84−5; Brandom, 2019, pp. 705-6).

Brandom's social-historical model of normativity has the potential to be used to clarify the conditions under which people are justified vouching for, and/or being held responsible for, the "testimony" of machines. Whether we are entitled to testify on the part of machines – and thus whether other people are justified in relying on them – depends, we submit, on the way in which that AI is deployed within social space, i.e., the social relations establishing norms of AI use. Rational standards for "trust" in machines are determined by the community as a whole. That is, if an appropriately situated, appropriately qualified, and rational agent is willing to vouch for a machine – to claim that its outputs give us reason to form beliefs about certain matters – then the larger community may judge them to be justified in doing so. In most circumstances, *part* of what is required of the voucher is that they have some way of assessing whether a machine is working properly, that is, producing outputs that correspond with some state of affairs. However, what it means to be appropriately qualified and appropriately situated is not a *purely* technological matter and does not admit of a purely technological specification. Rather, the capacity of the "voucher" to participate in practises of reason giving, which establish their own reliability as well

as the legitimacy of any further justifications they can provide for their own judgment in relation to the outputs of the AI (a certain amount of circularity is inevitable here), is the key consideration. This, in turn, will usually be a matter of the voucher's place in a broader network of social relationships, including, for instance, membership of professional associations or a history of interactions that establish a track record of sound judgment.

Even when the larger community judges that it is appropriate for an individual to vouch for a machine – that they are justified in doing so – there exists a further question about the nature, and range, of the inferences that are justified by the outputs – as vouched for by a human being – of machines. It is, moreover, a corollary of the Brandomian view being developed here that the testimony gap cannot be closed once-and-for-all by the collective determination of norms for legitimate epistemic reliance. It rather remains an ineliminable possibility that our collective willingness to allow a certain AI output as reason for belief will be found to be epistemically irresponsible. The socially determined entitlement to rely on such outputs might thus always be revoked or substantially revised, in just the same way that judge-made law might always revoke or revise previous decisions. No "story" a particular community has come to tell about what is right and wrong can be "final" (Brandom, 2019, p. 607). The testimony gap can never be closed or wholly filled, but only plastered over.

6 Conclusion

Short of abandoning technological society, we are doomed to live with the testimony gap. Acknowledging this fact, though, has the potential to open up possibilities for the form this society takes by expanding the range of considerations that determine when we will accept the "testimony" of machines.

At present, much of the discussion about the legitimacy of the outputs of AIs turns on the transparency and explicability of their internal processes. This is of course understandable given long-standing democratic commitments to transparency (Ananny & Crawford, 2016), and the assumption that engineers and managers *ought* to be in a position to understand and articulate the processes that lead to any given output, even if the challenges in meeting this obligation are well documented (Burrell, 2016). Our argument above might be taken to reinforce this approach: those who "vouch" for the epistemic reliability of some machine output *ought* to be able to provide a justification for their commitment to that machine's reliability that others can understand and assess; if they cannot do so, then the uptake of outputs from that machine is unjustified. However, justification for the take up of a machine's outputs into the space of reasons need not in all cases require a complete explanation of the machine's workings; it may, as we suggest above, be sufficient that the machine has been observed over time to be reliable, by a human agent taken by relevant peers to have a record of sound judgment. At the same time, however, crucial for the process of normative self-determination described above is that members of a given recognitive community who may rely on an AI system are in a position to understand and normatively assess the inferences and implications a machine is taken to license – or

put in classically pragmatist terms, that they are in a position to assess its *use*.[8] This assessment may involve ethical as well as pragmatic and epistemic considerations. To take the much-discussed example of the use of machine learning in "predictive policing" (Pease & McDaniel, 2021), if the outputs of a machine system are taken by a police force to justify the arrest of members of a historically over-policed and vulnerable minority community, continued use of that system is not *only* a question for those experts who are in a position to testify that the machine is working "reliably," but requires the making explicit of historical justifications regarding the institution of policing and its current social functions. That is, the justifications for the use of the system's outputs will also turn on the implications of that use, both cognitive and practical.

More broadly still, a further benefit of the Brandomian framework we have introduced here is its potential to remind us of historicity of our current moment, to self-consciously recognise our time as volatile and precarious. While we have reconstructed Brandom's argument that all judgments are implicitly constitutively contestable, the point, as always in Brandom, is to make the implicit explicit. We arguably live in a technological *Sattelzeit*, a transitional period in which technology and the norms governing it are in rapid development, and in which we anticipate a still inchoate future very different from our own.[9] It is crucial for those working on the ethics of new technology to cultivate a self-conscious relationship to the precarity and contestability of normativity, rather than to arrogate to themselves a one-sided responsibility to settle ethical questions once and for all. We ought rather to develop the "Spirit of Trust" that Brandom has sought to reconstruct in his most recent work: judging and acting in knowledge of the constitutive defeasibility of our particular commitments, trusting that other, themselves defeasible, agents will meet their obligation to contest and revise our claims. To flourish in an age of intelligent machines, we must ensure that we are ever conscious of the testimony gap and its implications for the epistemic foundations of our own beliefs. We can only rely on machines insofar as we trust each other.

Acknowledgements We would like to thank: Neil Levy for helpful comments on a draft of this manuscript; Joshua Hatherley for advice on the literature on medical AI; and, Peter Asaro, Lucy Suchman, and Elke Schwarz for advice on the literature on autonomous weapon systems. Professor Sparrow is an Associate Investigator in the Australian Research Council Centre of Excellence for Automated Decision-making and Society (Grant number CE200100005) and worked on this paper in this role.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. No funding was received to assist with the preparation of this manuscript.

[8] Vredenburgh (2022) makes a similar point: crucial to any institution's legitimacy is the degree to which individuals can identify with and affirm that institution's social role, which requires that individuals can assess the "normative character" of that institution. In the case of AI systems, "technical opacity" does not necessarily undermine the legitimacy of a given system, provided its users are positioned to normatively assess "the function it plays in the social world" (85).

[9] The connection between the so-called *Sattelzeit* in European history between 1750 and 1850 and our historical present is briefly noted by Nowotny (2021). There is much discussion about whether the so-called "fourth industrial revolution" is qualitatively continuous with earlier technological changes or promises a qualitative "leap" from modern industrial society to some genuinely novel social form (Brynjolfsson & McAfee, 2014; Srnicek & Williams, 2015). For a strident version of the latter position, see Floridi, 2014.

Declarations

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*, *20*(3), 973–989.

Asaro, P. M. (2012). A body to kick, but still no soul to damn: Legal perspectives on robotics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 169–186). MIT Press.

Brandom, R. B. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press.

Brandom, R. B. (2000). *Articulating reasons: An introduction to inferentialism*. Harvard University Press.

Brandom, R. B. (2002). *Tales of the mighty dead: Historical essays in the metaphysics of intentionality*. Harvard University Press.

Brandom, R. B. (2009). *Reason in philosophy: Animating ideas*. Belknap.

Brandom, R. B. (2019). *A spirit of trust: A reading of Hegel's phenomenology of spirit*. Belknap.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues* (pp. 63–74). John Benjamins Publishing Company.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1). https://doi.org/10.1177/2053951715622512

Cappelen, H., & Dever, J. (2021). *Making AI intelligible: Philosophical foundations*. Oxford University Press.

Cockburn, D. (1985). The mind, the brain and the face. *Philosophy*, *60*(234), 477–493.

Cockburn, D. (1990). An attitude towards a soul. In D. Cockburn (Ed.), *Other human beings* (pp. 3–12). Palgrave Macmillan.

Cockburn, D. (1994). Human beings and giant squids. *Philosophy*, *69*(268), 135–150.

Cockburn, D. (2021). *Wittgenstein, human beings and conversation*. Anthem.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. OUP.

Gaita, R. (1989). The personal in ethics. In D. Z. Phillips, & P. Winch (Eds.), *Wittgenstein: Attention to particulars* (pp 124–50). MacMillan.

Gaita, R. (1999). *A common humanity: Thinking about love and truth and justice*. Text Publishing.

Gaita, R. (2004). *Good and evil: An absolute conception* (2nd ed.). MacMillan.

Gleeson, A. (2001). Animal animation. *Philosophia*, *28*(1–4), 137–169.

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist*, *102*(2), 259–275.

Heinrichs, B., & Knell, S. (2021). Aliens in the space of reasons? On the interaction between humans and artificial intelligent agents. *Philosophy & Technology*, *34*(4), 1569–1580.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy artificial intelligence*. European Commission.

IPSOS (2021). *Global survey highlights continued opposition to fully autonomous weapons*. IPSOS. Retrieved February 3, 2024, from https://www.ipsos.com/en-us/global-survey-highlights-continued-opposition-fully-autonomous-weapons

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, *8*(4), 195–204.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183.

McDowell, J. (1996). *Mind and world*. Harvard University Press.

Moran, R. (2006). Getting told and being believed. In J. Lackey, & E. Sosa (Eds.), *The epistemology of testimony* (pp. 272–306). Oxford University Press.

Nowotny, H. (2021). *In AI we trust: Power, illusion and control of predictive algorithms*. Wiley.

Open Roboethics Initiative (2015). *The ethics and governance of lethal autonomous weapons systems: An international public opinion poll*. Open Roboethics Initiative. Retrieved February 3, 2024, from https://openroboethics.org//wp-content/uploads/2015/11/ORi_LAWS2015.pdf

Pasquale, F. (2015). *The black box society*. Harvard University Press.

Pease, K., & McDaniel, J. L. M. (Eds.). (2021). *Predictive policing and artificial intelligence*. Routledge,.

Pianalto, M. (2011). Speaking for oneself: Wittgenstein on ethics. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, *54*(3), 252–276.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, *28*(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

Redding, P. (2007). *Analytic philosophy and the return of Hegelian thought*. Cambridge University Press.

Renic, N., & Schwarz, E. (2023). Crimes of dispassion: Autonomous weapons and the moral challenge of systematic killing. *Ethics & International Affairs*, *37*(3), 321–343.

Sellars, W. (1956/1997). *Empiricism and the philosophy of mind*. Harvard University Press.

Schwarz, E. (2021). Autonomous weapons systems, artificial intelligence, and the problem of meaningful human control. *Philosophical Journal of Conflict and Violence*, *V*(1), 53–72.

Smith, H. (2021). Clinical AI: Opacity, accountability, responsibility and liability. *AI & Society*, *36*, 535–545. https://doi.org/10.1007/s00146-020-01019-6

Sparrow, R. (2004). The Turing triage test. *Ethics and Information Technology*, *6*(4), 203–213.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Sparrow, R. (2021). Why machines cannot be moral. *AI & Society*, *36*(3), 685–693.

Sparrow, R., & Henschke, A. (2023). Minotaurs, not centaurs: The future of manned-unmanned teaming. *Parameters*, *53*(1), Article14. https://press.armywarcollege.edu/parameters/vol53/iss1/14

Srnicek, N., & Williams, A. (2015). *Inventing the future: Postcapitalism and a world without work*. Verso.

Stop Killer Robots Coalition (2024). *Less autonomy: More humanity.* Stop killer robots. Retrieved February 3, 2024, from https://www.stopkillerrobots.org/

Taylor, C. (2014). Moral thought and ethical individuality. In C. Taylor, & M. Graefe (Eds.), *A sense for humanity: The ethical thought of Raimond Gaita* (pp. 141–151). Monash University Publishing.

Umbrello, S. (2022). *Designed for death: Controlling killer robots*. Trivent Publishing.

Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. *AI & Society*, *Online first*, 16April2021. https://doi.org/10.1007/s00146-021-01189-x

Vredenburgh, K. (2022). Freedom at work: Understanding, alienation, and the AI-driven workplace. *Canadian Journal of Philosophy*, *52*(1), 78–92.

Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical considerations of using ChatGPT in health care. *Journal of Medical Internet Research*, *25*, e48009.

Winch, P. (1980). The presidential address: 'Eine Einstellung Zur Seele'. *Proceedings of the Aristotelian Society*, *81*, 1–15.

Wittgenstein, L. (1989). *Philosophical investigations* (3rd ed.). Basil Blackwell.