Synthese (2021) 199:7061–7079
https://doi.org/10.1007/s11229-021-03105-8

First-person representations and responsible agency in AI

Miguel Ángel Sebastián[1] · **Fernando Rudy-Hiller**[1]

Received: 9 October 2020 / Accepted: 27 February 2021 / Published online: 19 March 2021

Abstract
In this paper we investigate which of the main conditions proposed in the moral responsibility literature are the ones that spell trouble for the idea that Artificial Intelligence Systems (AISs) could ever be full-fledged responsible agents. After arguing that the standard construals of the control and epistemic conditions don't impose any in-principle barrier to AISs being responsible agents, we identify the requirement that responsible agents must be aware of their own actions as the main locus of resistance to attribute that kind of agency to AISs. This is because this type of awareness is thought to involve first-person or *de se* representations, which, in turn, are usually assumed to involve some form of consciousness. We clarify what this widespread assumption involves and conclude that the possibility of AISs' moral responsibility hinges on what the correct theory of *de se* representations ultimately turns out to be.

Keywords Moral agency · Moral responsibility · Artificial intelligence · De Se representation · First-person representation · Consciousness

1 Introduction

There has been an impressive development in the field of Artificial Intelligence in recent years. Artificial intelligence systems (AISs) control trading in the stock market (Trippi & Turban, 1992) and make decision in surgical military strikes (Krishnan, 2009). They can drive our cars (Daily et al., 2017) and determine the music we want to hear or the products we should buy according to our interests (Linden et al., 2003). They also play a significant role in reaching medical diagnoses in certain cases (Kononenko, 2001). These are all contexts governed by moral—as

This is a fully collaborative paper. Authors appear in random order.

✉ Miguel Ángel Sebastián
msebastian@gmail.com

[1] Instituto de Investigaciones Filosóficas, UNAM, Universidad Autónoma de México, Circuito Mtro Mario de la Cueva, Ciudad Universitaria, Del. Coyoacán, 04510 Mexico City, Mexico

well as legal—norms, and actions performed in them are typically evaluated as correct/good/right or incorrect/bad/wrong. Hence, it is not surprising that the interest in ethical questions prompted by the possibility of AISs being agents in their own right—what is known as "machine ethics"—has increased exponentially as AISs become more integrated in our daily life (Müller, 2020; Wallach & Allen, 2010).

In particular, several recent writers have focused their attention on the question whether artificial systems can be held morally responsible for their activity (Coeckelbergh, 2020; Parthemore & Withby, 2014; Stahl, 2006; Sullins, 2006). Often, when the agent of a morally significant action is an adult human being, we hold them responsible for the action itself and its outcomes. On the other hand, we tend to resist (non-metaphorical) attributions of responsibility to artificial systems; we rather blame the creators of the system than the system itself. Imagine that you want to invest your saving and you hire a stock broker to trade actions. As a result of the broker's decisions you lose your savings. Without eluding your own responsibility for picking the broker, you consider them responsible for the result of their decision. Now imagine that the trading decision were made by an AIS. You will be more likely to blame the programmer than the AIS for the result. And this might be correct. However, recent improvements in learning mechanisms—e.g. AlphaGo Zero, (Silver et al., 2017)—together with the increasing complexity of AISs limit the programmer's control over the system's actions and outcomes: programmers do not know which rules the system is following to make its decisions. This does not mean that the programmer cannot be held responsible at all; however, the distance between the programmer's and the system's actions—the fact that the former does not seem to fully control the latter's actions—make it urgent to ask whether artificial systems can also be held responsible and, if so, to what extent (Sparrow, 2007).

Importantly, knowing these facts regarding the impressive advancements in machine learning does not necessarily modify the initial intuition that only human beings can rightly be held responsible and that attributions of responsibility to artificial systems are at best metaphorical. But what exactly supports this intuition, that is, why do we tend to resist attributions of responsibility to artificial systems? Should we abandon it or is it, at least *prima facie,* justified? More to the point, can AISs be held morally responsible for their actions and outcomes? The paper addresses this question. In particular, our goals are, first, to clarify the conditions that an AIS would have to meet in order to be appropriate in principle to hold it responsible for its actions and, second, to identify which of these conditions is the one that most likely produces resistance to the idea that such a system can really be a responsible agent. We will argue that the key element generating skepticism about the responsibility of artificial systems is the fact that one or more of the conditions for responsibility entail that responsible agents must possess first-person or *de se* representations, coupled with the assumption that this kind of representations necessarily entails consciousness.

To avoid misunderstandings, it's important to emphasize that this paper is not an exercise in "futurology", that is, we are not in the business of offering predictions about whether or which AISs would in fact be full-fledged responsible agents. Rather, ours is a conceptual investigation into the conditions for responsibility with an eye on establishing which of these conditions are the ones that may spell trouble

for the possibility of AISs being truly responsible agents. We leave it fully open whether current or future AISs can in fact fulfill those conditions because, as will become apparent, establishing this point requires first to develop a comprehensive and satisfactory theory of de se representations—something we still do not have. At the same time, however, we insist that our discussion is not merely a conceptual exercise but is also *about* AISs in the following sense: by clarifying the conditions that any AIS would have to meet in order to be a responsible agent, we are offering a substantive contribution to the philosophy of AI, even though we eschew predictions and avoid discussing specific AISs.

This is the plan of the paper. We offer, first, an overview of the conditions for responsible agency suited to the present topic. This will allow us to clarify the conditions that an agent would have to meet in order to be appropriate in principle to hold it responsible for its actions (Sect. 2). We then identify the capacities required to meet these conditions and discuss whether there is any reason to doubt that an AIS might possess any of them (Sect. 3). We identify the capacity to entertain first-person representations as the key component that could justify resistance to attribute moral responsibility to artificial systems. Finally, we address some objection to our argument (Sect. 4) and conclude with a brief summary discussing the possibility of having full-fledged responsible AISs (Sect. 5).

2 The conditions for responsibility

The distinctive mark of responsible agents is that they are accountable or answerable for their conduct. This means that they can properly be subject to a wide range of evaluations and reactions in response to the moral valence—positive or negative—of their actions and their consequences. The evaluations and reactions in question are *backward-looking* in that they are primarily concerned with assessing and reacting to an agent on account of her past conduct rather than with achieving some specific aim with them, such as improving the agent's behavior.[1] They also are *desert-entailing* in the sense that they represent their targets as deserving certain things just because they have performed the action in question, again irrespectively of whether this may bring good consequences for the agent or the wider moral community. What is thus deserved includes, at the very least, praise and blame, which are widely understood in terms of what Strawson (1962/2003) called "the reactive attitudes": pride, gratitude, and admiration in the case of praise; guilt, resentment, and indignation in the case of blame. On many accounts, what is deserved also includes formal and informal sanctions and rewards and, at the limit, legal punishment. Finally, the evaluations and reactions characteristic of responsibility are *quality-of-will tracking* because they are ultimately concerned not with overt behavior but with the inner

[1] This doesn't exclude the possibility that the evaluations and reactions characteristic of responsibility may bring (as they often do) beneficial side-effects; the point is just that they are not made *because* they bring positive consequences and also that the latter are not part of their appropriateness conditions (Pereboom 2014).

moral dispositions that such behavior reveals—what Strawson baptized the agent's "quality of will." So while it is true that when we hold someone responsible it is always on account of some action of them, it is the *agent* herself we are evaluating and reacting to and, therefore, we are interested in whether in acting as she did she displayed proper moral regard for others or not.

Given this focus on inner dispositions rather than on (merely) the causes of certain outcomes, attributions of moral responsibility presuppose certain capacities on the part of responsible agents. It is standard practice to group these capacities in two distinct conditions that are individually necessary and jointly sufficient for responsibility: a *control condition* and an *epistemic condition*. We will review them in turn.

2.1 Control condition

The condition that traditionally has gathered the lion's share of attention in the philosophical literature on responsibility is the *control condition*. It is concerned with whether the agent was sufficiently in control of her action so as to be true that it was up to her to perform it or not. Since the traditional assumption has been that this kind of control requires that the agent acted out of her own free will, it is also known as the *freedom condition*. It is crucial *not* to assume, however, that the freedom in question is necessarily incompatible with determinism and, more generally, with a naturalistic picture of the world. While there are still a number of philosophers who argue in favor of an incompatibilist conception of free will—although not necessarily a non-naturalist one—, the compatibilist conception is no less popular. In order to avoid prejudging our discussion from the start we will assume that a compatibilist conception is viable, since although it is hotly contested whether human agents possess a libertarian (i.e., incompatibilist) free will, as far as we know no one has ever suggested that artificial systems might. So if AISs turn out to be capable of possessing responsibility-relevant control, it most likely will be of a compatibilist sort.

According to the most popular compatibilist view of free will currently on offer, the control or freedom required for responsibility must be understood in terms of *reasons-responsiveness* (Fischer & Ravizza, 1998; Sartorio, 2016; Vargas, 2013). This means, roughly, that an agent acted freely if and only if she was at the time of action suitably receptive and reactive to the available reasons (including moral ones) favoring and disfavoring her conduct. The receptivity and reactivity in question are cashed out in terms of a suite of cognitive and executive capacities allowing the agent to detect normatively relevant features of her environment—such as that a person needs help—and respond appropriately to them—provide the needed help.

Importantly, many authors acknowledge that the requisite responsiveness can come about *without* the involvement of conscious and explicit deliberation, since habitual and automatic actions can, and often do, count as exercises of the relevant capacities (Fischer & Ravizza, 1998, p. 86; Vargas, 2013, p. 217). Thus, for example, when you automatically dodge a frisbee headed in your direction you manifest appropriate responsiveness to reasons, as you do when during your habitual drive to work you monitor the road for obstacles while thinking about something else.

Automatic responsiveness can also be manifested in morally loaded contexts, as when a person who sees a drowning child jumps into the water without giving it a thought.[2]

Finally, it is worthwhile to make it explicit why this characterization of free will and control is both compatibilist and naturalist. It is compatibilist because it is close to uncontroversial that even in a deterministic universe people could manifest suitable responsiveness to reasons, a thesis that even those philosophers who argue that reasons-responsiveness is *in*sufficient for responsibility accept (Pereboom, 2014). And it is naturalist because the capacities invoked by the account are consistent with a fully naturalistic picture of the world, according to which everything happens in accordance with, and is explainable by, natural laws, and in which these capacities are realized in wholly material systems subject to those same laws.

2.2 Epistemic condition

Even though responsible agency seems to be compatible with automatic and habitual responsiveness to reasons and can therefore dispense with conscious deliberation (at least sometimes), many philosophers think that it cannot dispense with *awareness* altogether.[3] This is because, in order for an agent to control her actions in a robust sense, she must be aware of certain things. The condition on responsibility that deals with the sort of awareness that is thus required is known as the *epistemic condition*. There is controversy regarding whether it is really an independent condition or whether it is merely an aspect of the control condition (Levy, 2011; Mele, 2010, Ch. 5). As we just hinted we favor the latter view, since the control condition, when understood in terms of reasons-responsiveness, already incorporates a dimension related to the agent's capacity for becoming aware of normatively relevant features of her situation. However, for expository purposes we find it convenient to characterize the awareness required for responsibility on a separate basis, with an eye on our larger goal of identifying which capacities of responsible agents are supposed to be inaccessible to artificial systems.

The epistemic condition is concerned with two main questions. First, of *what* must an agent be aware in order to be responsible for a particular action or outcome? Second, in *what way* must she be aware of it? The first question is about the *content* of the requisite awareness, while the second is about the *kind* of awareness that is required (Rudy-Hiller, 2018). Let's focus on the first question. There is general agreement that the contents of awareness that are relevant for responsibility include awareness of the *action* the agent is performing; of its *moral valence* (that the action

[2] See also Shepherd (2014) for a theory of control in the philosophy of action that is compatible with the possibility of control being exercised without conscious deliberation. Of course, much more can be said in this regard, but we can't do so here given the paper's scope.

[3] Sher (2009) defends the possibility of "responsibility without awareness" (this is his book's subtitle), but he argues against a very particular understanding of awareness as involving occurrent beliefs. As we will see below, there are other, and more plausible, conceptions of awareness available.

is good, right, or permissible, or bad, wrong, or impermissible); of its *consequences*; and of the *alternative* courses of action open to the agent.

Several clarifications are in order. First, awareness of action must occur, as it is often put, "under an appropriate description" (Levy, 2014, p. 37), that is, a description under which the target action come off as intentional. So if, for instance, Brian is aware merely of the fact that he is contracting his index finger, then he is not aware under an appropriate description of his action of pulling the trigger and, consequently, the latter is not intentionally performed. Importantly for our topic, being aware of an action under an appropriate description includes awareness of *oneself* as the one performing the action in question. If one lacks this awareness, the target action does not count as intentional either. For example, if to the question "Why are you ringing that bell?" one answers "Good heavens! I didn't know *I* was ringing it", this shows that the action of ringing the bell was not intentionally performed (Anscombe, 1963, p. 51).[4] This will be relevant below because awareness of oneself as the author of an action entails the possession of first-person representations.

Second, awareness of the action's consequences obviously does not need to extend to *all* imaginable consequences, which may well be impossible for a finite agent to compute. The usual rule of thumb that helps determine which consequences are relevant is the "reasonable person" standard taken from tort law, according to which an outcome is morally relevant if a reasonable person in the agent's situation would have foreseen it when deliberating about what to do. Also, it is enough for satisfying this requirement if the agent is aware of the relevant consequences under a suitable general description. Thus, if an action has as one of its consequences the harming of a specific individual in a particular way, the agent does *not* have to foresee this in order to be responsible for the outcome. Rather, it suffices if she is aware of the fact that her action poses a significant risk of harm to others (Fischer & Tognazzini, 2009).

Third and finally, a similar proviso about relevance applies to awareness of alternative courses of action. In order to be responsible for a wrong action, an agent need not be aware of *all* the available alternatives but only of at least one *permissible* alternative. If, on the contrary, the agent reasonably believed that she had only one course of action available to her, then plausibly she is not responsible for it (Levy, 2011, p. 111).[5]

We will conclude this section by briefly exploring the second question broached above concerning the *kind* of awareness that is necessary for responsibility. First, we need to know *which* mental states constitute the relevant awareness. There is controversy about whether reasonable belief, justified belief, or belief that amounts to knowledge is what the epistemic condition demands. Everyone agrees, however, that *some* kind of belief is required, so for our purposes we will stick to this minimal

[4] And, according to many philosophers, this entails that one is not responsible for it unless one is culpable for one's ignorance about what one is doing (Smith, 1982; Zimmerman 1997).

[5] In the case of *right* actions, however, the agent does not need to be aware of alternatives in order to be responsible (i.e., praiseworthy) for performing such actions.

common ground and assume that belief simpliciter (or an equivalent representational state) constitutes the relevant awareness.[6]

Second, we need to know *how* these beliefs must be entertained. A usual distinction is drawn between *occurrent* and *dispositional* awareness. The former involves an occurrent conscious belief about the action one is performing, its moral valence, its possible consequences, and available alternatives to it, while the latter involves a dispositional or unconscious belief about the same. We find appealing Levy's account of the kind of dispositional awareness that is required for responsibility, which he defines in terms of *personal availability*: "Information is personally available … when the agent is able to effortlessly and easily retrieve it for use in reasoning and it is online" (2014, p. 33). On this definition, an agent satisfies the epistemic condition if, at the time of deliberation (if there is such) and action, she *can* (but not necessarily will) recall information about the contents mentioned above without having to make any special effort but rather when prompted to do so by a wide range of cues and such information actually plays a role in guiding her behavior (34), as when an absent-minded driver manages to successfully get home while being dispositionally aware of relevant information such as the position and speed of other vehicles, the presence of pedestrians, traffic signs, etc. As this example suggests, it is clear that something like Levy's account of the kind of awareness required for responsibility must be right, because otherwise we would be unable to attribute responsibility for negligent conduct to those drivers who, on account of their distraction, unintentionally harm others. Also, a dispositional account of awareness is needed to accommodate responsibility for habitual conduct and to Block the possibility of dodging responsibility by engaging in self-deception (Haji, 1997, pp. 537–539).

3 Capacities required for the conditions

In the previous section we have presented the key conditions on which the attribution of responsible agency depends according to the relevant philosophical literature. In particular, we have remarked that there are two (perhaps interdependent) conditions that a system has to satisfy in order to be considered a responsible agent: the control condition and the epistemic condition. In this section we will investigate the cognitive requirements that a system has to fulfill in order to satisfy these conditions. The aim is to understand the source and rationale of the resistance to think that artificial systems can be full-fledged responsible agents—see Sparrow (2007) for a representative skeptic about the possibility of AISs being truly responsible agents.

[6] Those who think that mere belief is insufficient for responsibility stress the intuition that something not as demanding as correctness, truth or accuracy but more demanding that mere representation is required for the attribution of responsibility. This is what the property of being 'justified' is supposed to do. An analysis of the different views about epistemic justification is beyond the scope of the paper, but we can consider for current purposes a majoritarian view that roughly holds that a belief that p is justified for an agent, S, if and only if forming the belief that p is permitted for S; that is, S is not required not to believe that p (e.g., Echeverri, 2019; Goldman, 1986, p. 59; Littlejohn, 2012, p. 8; Pollock & Cruz, 1999, p. 123; Silva, 2017; Wedgwood, 2012, p. 274).

3.1 Control condition

The control condition for responsible agency demands that the system has certain cognitive and executive capacities that allow it a) to detect the normatively relevant features of her environment and b) to respond appropriately to them. Imagine that a person is asking for help as she is drowning. The first condition impose that a system cannot be held responsible for its actions in these situations unless it is sensitive—i.e., it has the capacity to detect—the morally relevant features of the situation—that someone needs help. The second condition imposes that the system has the capacity to respond appropriately to this situation. An adequate response in this case could be something like the capacity to rescue the person—maybe without putting its own existence in risk—or to call for help. Current artificial systems do have the required executive abilities to respond appropriately in at least some morally relevant cases demand and hence condition (b) provides no reason to cast doubt on the idea that AISs can be responsible agents. As with regard to condition (a), one might think that an AIS might be unable to detect normatively relevant features because it cannot have detectors for normatively relevant features. We cannot figure out what would be the motivation for such a claim. Fortunately, our imaginative capacities in this regard are irrelevant because the idea is simply misguided. In general, detecting a state or event S does not require that one has some sort of specific detectors for S; it rather requires a distinctive state of the system, R, that correlates with the environmental condition S.[7] For example, in vision we can detect apples despite the fact that we only have specific detectors for basic properties like shades, colors, etc. Likewise, we seem to lack specific detectors for people needing help and we detect the normatively relevant facts by detecting, for example, that they are yelling and moving their arms—which we might detect in turn by detecting more basic features. AISs have no problem to detect those properties on the basis of which we detect the normatively relevant properties.

One might reasonably remark that being able to detect the very same properties that we detect might not be sufficient for detecting the normatively relevant facts. For example, one might think that detecting the normatively relevant facts requires, on top of the capacity to detect certain aspects of the situation—a capacity that AISs might share with us—, to employ such information to become aware of the reasons to act. So, although nothing prevent the use the available information for further tasks, there is room for reasonably calling into question that AISs have the capacity to detect normatively relevant facts for reasons linked to the awareness of the reasons to act. This brings us straightforwardly into the analysis of the epistemic condition.

[7] This requires that the presence of R changes the probability of S ($P(S|R) \neq P(S)$). This is not sufficient for R to represent S—for discussion see Artiga and Sebastián (2020). As presented, the control condition requires detection rather than representation. We deal with representational properties when considering the epistemic condition below.

3.2 Epistemic condition

On the construal we favor, the epistemic condition on responsible agency shows that responsibility-relevant control requires that the agent is aware of certain things. As we explained above, this requires an elucidation of both the content and kind of awareness that is relevant for responsibility attributions. We turn now to an investigation of the cognitive capacities that are involved in securing the relevant awareness.

3.2.1 The kind of awareness

First, we need to know what kind of mental states constitutes the relevant awareness. As we noted above, the state in question is typically characterized as belief. Roughly, to believe that *such-and-such is the case* is to take it to be true that *such-and-such is the case*—to take it that the state of affairs described by the sentence "*such-and-such*" obtains.[8] The analysis of different attitudes is typically given in functional terms and there is no reason to think that—leaving consciousness aside—an artificial system cannot satisfy the characteristic function of beliefs. In particular, the attitude of taking *p* to be true doesn't entail anything beyond the capacity to draw certain inferences from *p*, the disposition to sincerely report that *p* and the disposition of letting action be guided by *p*.

We can safely accept the truism that one can belief that *p* is the case only if one can represent that *p*—only if one can be in a state that represents *p*. One might thus think that the reason why an artificial system cannot be a responsible agent is precisely that artificial systems cannot have genuine representations. For example, according to John Searle (1990) genuine representation depends on consciousness and most people are reluctant to attribute conscious experiences to AISs. According to Searle, artificial systems can be said to hold representational states only derivately, since in their case those states depend on the content of the representations of their conscious creators, and one might argue that responsible agency requires genuine representations.

However, this view is not very popular nowadays. Against it, naturalistic theories of mental content attempt to explain what it takes for a system to entertain representational states in non-intentional terms. These theories exploit different causal (Fodor, 1987; Kripke, 1980; Putnam, 1981), functional (Millikan, 1984, 1989; Neander, 1991, 2017; Papineau, 1984) or structural relations (Block, 1986; Cummins, 1996; Watson, 1995) to offer an account of representation. On the basis of such accounts, most theorists are willing to attribute to artificial systems the same kind of representational states that we attribute to our own cognitive systems. Recent

[8] It is worth stressing that this does not require that the subject explicitly believes *that such-and-such is the case is true, which is a* different and higher order belief—the belief that *such-and-such is the case*, for instance, doesn't require possession of the concept of "truth". Moreover, as we have seen, some authors have insisted that the epistemic condition requires a particular kind of belief, like *justified* belief. If AISs can entertain representational belief-like states (as they surely do) then they can also entertain *justified* belief-like states according to majoritarian views of justification—see fn. 6 above.

research on the notion of representation as it is used in cognitive science—which contributes to explain why animals succeed or fail in various goal-directed activities (Shea, 2018)—can help us to illustrate the absence of significant differences between human and non-human representations. This supports the claim that AISs can have representational states.

Explanations in terms of representations allow us to make sense of successes and failures of goal-directed behaviour. These sort of behaviours or outputs of a system are called by Shea the *task functions* of the system. According to Shea, for the output of a system (like movement or behaviour) to be explainable in representational terms, the system has to be able to produce this output in response to a variety of inputs and across an array of external conditions as a result of a stabilizing process, thereby exhibiting a teleological function. Shea embraces the etiological view that a behavioural trait has a function in virtue of its history, where natural selection is the typical stabilizing force that fixes the function of the output. But, as he notices, biological evolution is not the only stabilizing force. There are other forces operating at the level of individuals rather than populations that can also play the required role in fixing the function of a trait or output, for example feedback-based learning and contribution to persistence (ibid., ch. 3).

Representations are internal states (vehicles) that bear exploitable relations to distal states of the world, where exploitable relations are informational correlations between vehicles and the distal states in the world, or structural correspondences between vehicles in the system and relations in the world—as it happens in the case of a map, where there is a structural correspondence between the features of the vehicle, the map, and what it represents. Representations play a role in the explanation of an output when the system performs the function through the computation on these internal vehicles that exploits the relations they bear to the world. AISs have then task functions, because their outputs can be stabilized and because they have internal states that stand in relations to the states of the world that can be exploited in the computations that give rise to the corresponding task function. Therefore, the sort of intentional explanation of behavior that is adequate in our case—and in the case of non-human animals—is also adequate in the case of AISs, so there is no reason to think that AISs cannot have representations thus understood.

Another important source of resistance to the idea that AISs can genuinely be morally responsible concerns consciousness (Torrance, 2012). As we noted in Sect. 2, most philosophers think that responsible agency requires awareness of some kind or another, and awareness is usually equated with consciousness. The problem here is that many people tend to resist the idea that AISs can have conscious experiences at all (Gray & Gray, 2007).

The terms ‘consciousness’ is used to mean different things, some of them closely related. For current purposes, it is important to focus on the distinction that Ned Block (1995) draws between access and phenomenal consciousness. Access consciousness is related to the access the subject has to the information carried by a certain state. A mental state is access-conscious if and only if, roughly, the content of the state is available for belief formation and rational control of action. In contrast, the term ‘phenomenal consciousness’ is used to refer to our subjective experience. We can say, using Nagel’s (1974) expression, that a mental state is phenomenally

conscious if and only if there is something it is like to be in that state. The conceptual distinction is clear, but the conclusions to be derived from it have remained controversial since the publication of Block's paper. Do these concepts correspond to different properties? In other words, is there access consciousness without phenomenal consciousness or phenomenal consciousness without access? In the search for an answer to this question, the debate has recently moved away from the conceptual to the empirical domain, focusing on the possibility of phenomenology without access. The notion of access consciousness has been refined to that of *Cognitive Access* and the question now is whether the neural basis of phenomenal consciousness can be disentangled "from the neural machinery of the cognitive access that underlies reports of phenomenal consciousness" (Block, 2007, p. 399). This intends to answer the question whether cognitive systems like us can have phenomenally consciousness states in the absence of cognitive access to the content of these states, which is a crucial question for the scientific study of consciousness.

We contend that the relevant notion of consciousness for the attribution of responsible agency is that linked to information: what is relevant is that the pertinent information—which we will analyze below—is available to guide action. Levy (2014, ch. 3) argues in detail that the notion of consciousness required for moral agency is closely relate to that of access consciousness.[9] In particular, as we have seen, he claims that what is required is that information is *personally available*, meaning that it can be retrieved effortlessly—in the absence of cues—by the agent for reasoning and that it is online—actually playing the role in guiding behavior. If, as we have seen, AISs can indeed have representations, and if they can perform morally relevant actions through the computation on these internal representation exploiting the relations they bear to the world, the information can be used in inferences and guide their behavior. So, in the relevant sense the pertinent information is accessible to them and it is personally available.

Further support for thinking that our notion of responsible agency is not necessarily linked to phenomenal consciousness is provided by a study conducted by Gray and Gray (2007). They show that our attributions of mentality have two dimensions: one related to agency and one related to subjective experiences. Human adults score high in both dimensions—we attribute to human adults both moral responsibility for their actions and subjective experiences—but they can be dissociated. For example, they show that people tend to attribute conscious experiences to "simple" animals like frogs without attributing agency to them. On the other hand, we tend to attribute absolute responsible agency to God but no subjective experience to them. On their part, robots score at the lowest level with regard to subjective experiences and moderately with regard to action (it would be interesting to see how these intuitions have changed since then with the development and integration of AISs in our life).

[9] According to Levy, there are subtle differences between is notion of personal availability and Block's notion of access and that it is an open empirical question whether information that is personally available to us is also available to broad variety of consuming systems (pp. 35–36). But he thinks that there are good reasons to answer in the affirmative. Be that as it may, AISs can have information personally available in Levy's sense and available to the sort of consuming systems required by Block.

3.2.2 The content of awareness

If we grant artificial systems the possibility of having representational states, the question now is whether those states can have the content required for moral responsibility. In Sect. 2, we identified four different contents that are plausibly involved in moral action: the agent has to be aware—or being in a position to be aware if a dispositional view is preferred—of the action that they perform, of its moral valence, of its consequences and of the alternative courses of action.

The last two types of content seem not to be a problem at all for AISs. Current artificial systems are able to predict the consequences of different courses of action—like buying or selling stocks—much better than any human agent. They can be aware of the consequences of the decisions they make—i.e., they can represent such consequences, like increasing benefits—and learn on their basis. Thus, there is no reason to doubt that AISs can entertain these types of contents.

What about the moral valence of actions? AISs are able to entertain explicit representations of the action's moral valence. Such a representation doesn't seem to require anything more than a function that maps pairs of sets of actions and circumstances into a particular moral valence, i.e., good/right/permissible or bad/wrong/impermissible. This can be implemented unproblematically in an artificial system. Moreover, with the adequate feedback, the artificial system can flexibly adjust such a mapping in a learning process following a Bayesian model.

The key condition is the awareness of our own actions. In order to be held responsible, one has to aware of the action one performs as an action that *one oneself performs*: it requires self-awareness. Consider generally the case of thought. There are several ways in which one can think of oneself. One can think of oneself appealing to some description, for example you think of the current reader of this paper. If one happen to be identical with a certain entity—maybe a certain body—, one can think of oneself by thinking of that entity. One can also think of oneself demonstratively (that person) or using their name. In all those cases, it seems that one can wonder whether one is really thinking of oneself, as o might ignore that one is a certain entity, that one is the person they are demonstrating, etc. On the contrary, one can think of oneself in a first-personal or *de se* way, in such a case there is no room for such a wonder (Castañeda, 1966; Chisholm, 1981; Lewis, 1979; Perry, 1979).

We have to be aware of our own action in a *de se* way. To illustrate, suppose that, after months of confinement, the medical authorities have developed a substance that changes the color of the nose only when the subject is infected by COVID-19. The substance is distributed within the population so that subjects who are infected remain home in order to avoid spreading the disease. Marta is aware of the consequences of going to the cinema with COVID-19, she believes that going to the cinema with COVID-19 is wrong and that people with COVID-19 should remain at home. Marta is fully aware of the moral valence of the different actions she can take, its consequences and alternatives. Marta takes the new substance, she looks carefully into the mirror and she does not detect any change in color; so she decides to go to the cinema. Unfortunately, a sudden change in illumination in her room was responsible for her misperception: her nose has changed colors and she has COVID-19. In these circumstances we would not find Marta responsible for her

action, because she is not aware of going out with COVID-19. Imagine now that, as she goes out, a neighbor sees her and shouts "Marta is going to the cinema with COVID-19". Marta is aware, after hearing the shout, that Marta is going to the cinema with COVID-19. However, this is not sufficient to find Marta responsible for going out with COVID-19. She needs to be aware of the action described as "Marta is going out with COVID-19" as her *own* action; she has to be aware that she herself is going out with COVID-19, something that she would convey with the expression "I am going to the cinema with COVID-19". If Marta isn't aware of the fact that she is the same Marta her neighbor is referring to, she can know "Marta is going to the cinema with COVID-19" without knowing "I am going to the cinema with COVID-19", and hence she would not be responsible.[10]

We speculate that the required first-personal awareness of action is the fundamental element behind the widespread reluctance to attribute moral agency to artificial systems. The reason is that it is not clear what is required from an artificial system to entertain this sort of first-personal awareness. Several authors within the debate have stressed the relevance of a concept of self for responsible agency,[11] but they have not pinned down the reasons for such a claim nor have they offered an adequate understanding of the form of self-awareness that is required. For example, Parthemore and Withby (2014) claim that an explicit concept of self is required. This is the concept of "who and what she thinks she is.[…] explicit concept of self-as-myself, as an intentional and distinctively cognitive entity" (147). It is not enough that the agent is aware of itself as performing an action, we need a *de se* awareness. This is not a matter of being aware of a particular entity as an intentional and distinctively cognitive entity. Marta might be aware that Marta is an intentional distinctively cognitive entity, but as the example above shows this is insufficient for the required *de se* awareness. Parthemore and Withby go on in an attempt to clarify the required notion and claim that "She must be able to hold herself responsible: and that she cannot do without full self-conscious awareness. She must, so to speak, be able to recognize herself in a mental mirror" (147). The metaphor is not very helpful either. Moreover, this suggests without further justification that consciousness is required and it is unclear that AISs can be consciously self-aware—if this is understood as involving phenomenal consciousness. Maybe they have the informational sense associated to access consciousness in mind. Then it would be a claim about the kind of access that the subjects need to have to the content of self-awareness—something that is not problematic for AISs as we have previously seen—, but it does not help to understand the required content. We have seen that what is required is *de se* content. The key question is: what is required for having this sort of content? The example above

[10] Several authors (Cappelen & Dever, 2013; Magidor, 2015) have argued that *de se* representation is not a especial kind of representation. If they are right and our reasoning in this paper is correct, then there are good reasons to think that our intuitions regarding attributions of moral agency to AISs are not justified.

[11] For an independent argument in favor of the need of first-person awareness on the basis of the need of autonomy see Neely (2014). See Sect. 4 for an objection to our argument based on the connection between autonomy and responsible agency.

illustrates that de se awareness is not a matter of having a particular variable, a tag or a name that is exclusively attached to the subject.[12]

Unlike information regarding moral valences, consequences and alternatives, first-person information is not the sort of information one can write down in a book. This is the sort of information that we attribute to self-conscious agents and the one that one may be reluctant to attribute to artificial systems. We can think of semantic information in terms the way it restricts the space of alternatives compatible with the truth of the information. If a system does not know anything about its environment and it acquires information regarding the environment, the amount of possibilities that are compatible with what it has learned decreases. For example, a scenario where there is a TV, a chair or a dinosaur on the right corner of the room is compatible with the information that there is sofa on the left corner, but not with there being a chair at the left corner: this is a possibility that the information that there is a sofa on the left corner rules out. With this idea in mind we can think of semantic information that an agent possesses as the set of situations or worlds that are compatible with the truth of the information that the agent possesses. Adapting Lewis (1979)'s two goods argument we can understand why this sort of information is insufficient to characterize first-person information. Imagine a certain environment E that contains two AISs, C3PO and R2D2, one of them naming out loud the normally relevant events of the environment and the other just beeping. Imagine that they both are omniscient in the sense above mentioned with regard to E: they both know exactly which environment they inhabit. In particular, they both know that R2D2 is beeping and C3PO is naming out loud. However, intuitively, from this knowledge they cannot obtain knowledge regarding which one of the two AISs they are. C3PO ignores the facts that it would express with the statements `I am naming out loud' or with `I am C3PO'. These facts are not entailed by any of the non-perspectival facts that C3PO already knows.

If you were in this environment and your cognitive capacities could be enhanced in such a way that you had the third-person knowledge the AISs have, I predict that you would have no problem knowing who you are. The knowledge that C3PO has together with our conscious experience puts us in a positions to have first-person knowledge. This does not logically entail that (phenomenal) consciousness is required for first-person knowledge. However, some authors have indeed appealed

[12] An anonymous referee has called our attention to the connection between first-personal information and the problem of personal identity. While it is indeed reasonable to think that the two are linked (for example, if personal identity depends on the continuity of mental states—some of which are first-personal), research on *de se* representations is typically conducted independently of concerns about personal identity. Furthermore, while there are interesting issues for responsibility derived from the problem of personal identity, usually derived from so-called "manipulation cases" (Mele, 2019), there is nothing we can think of in the case of AISs that can be particularly problematic in contrast to the case of human beings in this regard. Moreover, interesting as it might be for responsibility, continuity is not per se relevant for the attribution of full-fledged responsible agency, which is our focus in the paper. Personal identity concerns what makes a person S_1 the same person as S_2 (analogously for AISs). This is relevant if we want to punish S_2 for an action S_1 performed at an earlier time, but not to determine whether S_1 was blameworthy for performing it: an agent can be held responsible for her action at time *t* even if she ceased to exist for any time after *t*. We thank an anonymous referee for pressing this issue.

to our conscious experience to explain *de se* content (García-Carpintero, 2017; Peacocke, 2014; Recanati, 2007, 2012). On the other hand, some others attempt to explain in naturalistically acceptable terms *de se* representation without taking consciousness for granted (Bermúdez, 2016; Sebastián, 2012, 2018). Whether AISs can be responsible depends on these issues.

4 Objections and replies

We turn now to a couple of objections to our argument that the possibility of AISs' entertaining *de se* representations is the key elements fueling skepticism about their being genuinely responsible agents. The animating thought behind these objections is that, while *de se* representations may be important in determining who (or what) is a responsible agent and who (or what) isn't, the real source of resistance lies elsewhere, namely in the (im)possibility of AISs' being truly autonomous agents and their apparent inability to express moral meaning through their actions.

1. *Autonomy*. According to the autonomy objection, the real problem with the suggestion that AISs could be responsible agents is that AISs aren't the source of their decisions and actions; their designers are. Thus, they lack the sort of autonomy that ordinary human adults typically have and which, according to several theorists, is a necessary condition for moral responsibility (Haji, 1997; Kane, 1996). In response, it's noteworthy that this sort of "sourcehood objection" has been raised against the possibility of persons' being morally responsible as well, on the grounds that both determinism and indeterminism preclude the autonomy required for genuine responsibility (Pereboom, 2014). Thus, the worry that seemingly responsible agents aren't really the source of their decisions and actions and therefore lack the autonomy required for moral responsibility isn't restricted to AISs. Since in this paper we have left aside this kind of metaphysical skepticism about responsibility, we needn't worry about the prospect that AISs may not be autonomous or the source of their decisions and actions in a very strong sense (quite likely, we human beings aren't anyway). What is important for present purposes is that some AISs—as, for example, those that result from evolutionary algorithms (Vikhar, 2016)—do have a significant degree of flexibility and autonomy or leeway regarding their decisions and actions, since the latter aren't fully controlled or predetermined by their designers. We contend that this modest degree of autonomy is sufficient for responsibility.[13]
2. *Moral meaning*. A natural reaction to the argument rehearsed in the previous sections starts by conceding that some AISs can formally satisfy the control

[13] One way to put our argument in the paper is this: regarding the main conditions proposed in the moral responsibility literature, either AISs meet those conditions or human beings don't always meet them and yet we hold them responsible anyway—as it is the case with the sourcehood condition just discussed. Of course, our main point is that the key contentious condition that human beings do meet but AISs may not is the one related to awareness of one's own actions, which entails the possession of *de se* representations. We thank an anonymous referee for suggesting this wording of our position.

> and epistemic conditions on moral responsibility, and yet insists that they can't be responsible agents because they don't really understand *what* makes wrong actions wrong and, relatedly, they don't understand *why* certain kinds of being merit moral consideration. Thus, the objection goes, even though AISs can be programmed so as to be able to recognize certain actions as impermissible, they aren't really capable of expressing "moral meaning" through their conduct, where an action's moral meaning depends on whether the agent is capable of understanding how her behavior affects the moral relationships she can have with others (McKenna, 2012).

In response, it's important to note that the contention that responsible agency involves the ability to transmit moral meaning in conduct is pursued by analogy to the ability that competent speakers of natural languages possess to transmit linguistic meaning in their utterances (McKenna, 2012). Thus, we can address this objection indirectly by seeing whether sophisticated AISs are capable of transmitting linguistic meaning and, more generally, of sustaining an intelligible conversation. If they can, then they might well be able to sustain a "moral conversation" of the sort that characterizes responsible agents. And, from our point of view, it is hard to resist the idea that cutting-edge AISs such as GPT-3 (Generative Pre-trained Transformer 3, an autoregressive language model that uses deep learning mechanism to produce human-like text) already have the capacity to transmit linguistic meaning in conversation. GPT-3 is able to maintain intelligible conversations about almost any topic. The reader can search the Internet for some such interactions and judge by themselves. For illustration, one can read the opinion of several philosophers and GPT-3′s reply in *Daily Nous*[14] or GPT-3′s article in *The Guardian.*[15]

Therefore, if the metaphor of responsible agents' transmitting moral meaning through their conduct is pursued in analogy with their ability to transmit linguistic meaning, then at least some cutting-edge AISs may be fully capable of transmitting both.

5 Conclusion

Many philosophers are inclined to resist the attribution of moral responsibility to AISs. In this paper we have sought to bring light to this debate by clarifying first the conditions on responsible agency and then assessing whether there are good reasons to doubt that AISs could in principle satisfy them. We have argued that the only condition one might reasonably doubt that AISs could ever fulfill is the capacity for awareness of their own actions, given that the latter requires *de se* representations. Therefore, the answer to the question on whether AISs can be responsible agents depends on an account of what is required from a system to entertain *de se* representations. Such an account is still missing, and in the meantime we can

[14] https://dailynous.com/2020/07/30/philosophers-gpt-3/.

[15] https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

only speculate on the basis of available approaches. As we have seen, some authors (García-Carpintero, 2017; Peacocke, 2014; Recanati, 2007) have attempted to explain *de se* representation on the basis of consciousness. If they were in the right track, then we would not be in a position to evaluate whether AISs can be morally responsible because we are very far from being in a position to determine the conditions that a system has to satisfy to have conscious experiences. However, appealing to consciousness to explain *de se* representation is very controversial, because it attempts to explain something mysterious (*de se* representation) in terms of something that is even more mysterious (consciousness). For this reason, other authors have looked for alternative routes. For example, Sebastián (2012, 2018) grounds our capacity to have basic *de se* representational states on the role such states play in self-maintenance, endorsing the widely accepted view in biology that living organisms are self-maintaining systems, systems that favor the conditions for their own maintenance. On this approach, whether AISs can be moral agents depend on whether they can be self-maintaining systems. We cannot evaluate this possibility here, but we do not see any reason that speaks in principle against it. And indeed, the notion of self-maintenance—which can be traced back at least to Aristotle (Godfrey-Smith, 1994; McLaughlin, 2001)—has gained popularity in contemporary science thanks precisely to the work of cyberneticians (e.g. Ashby, 1947; von Foerster, 1960). But this is a topic for another occasion. Here we hope to have contributed to move the debate forward by showing where we have to look for in order to ascertain whether AISs could be truly responsible agents.[16]

Conflicts of interest None.

References

Anscombe, G. (1963). *Intention*. Harvard University Press.

Artiga, M., & Sebastián, M. Á. (2020). Informational theories of content and mental representation. *Review in Philosophy and Psychology, 11,* 613–627.

Ashby, W. R. (1947). Principles of the self-organizing dynamic system. *The Journal of General Psychology., 37*(2), 125–128.

Bermúdez, J. (2016). *Understanding "I": Language and thought*. Oxford University Pres.

Block, N. (1986). Advertisement for a semantics for psychology. In P. French, T. Uehling, & H. Wettstein (Eds.), *Midwest studies in philosophy* (Vol. 10, pp. 615–678). University of Minnesota Press.

Block, N. (1995–2002). On a confusion about the function of consciousness. In: Block, N. (Ed.) *Consciousness, function, and representation: Collected papers* (Vol. 1). Bradford Books.

Block, N. (2007). Overflow, access, and attention. *Behavioral and Brain Sciences, 30,* 530–542.

Castañeda, H.-N. (1966). 'he': A study in the logic of self-consciousness. *Ratio, 8,* 130–157.

Cappelen, H., & Dever, J. (2013). *The inessential indexical: On the philosophical insignificance of perspective and the first person*. Oxford University Press.

[16] An ancestor of this paper was presented at the AI symposium at the Canadian Philosophical Association Annual Congress 2018. We are grateful to the participants at this event, especially to Martin Gibert, Dominic Martin, Vincent Muller, Pierre Poirier and Christine Tappolet. Financial support for this research was provided by DGAPA projects IA400218 and IG400219, and the Spanish Ministry of Science, Innovation and Universities via research Grant PGC2018-095909-B-100.

Chisholm, R. M. (1981). *The first person: An essay on reference and intentionality*. University of Minnesota Press.
Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justifcation of explainability. *Science and Engineering Ethics, 26,* 2051–2068.
Cummins, R. (1996). *Representations, targets, and attitudes*. MIT Press.
Daily, M., Medasani, S., Behringer, R., & Trivedi, M. (2017). Self-driving cars. *Computer, 50*(12), 18–23.
Echeverri, S. (2019). Emotional justification. *Philosophy and Phenomenological Research*, *98*(3), 541–566.
Fischer, J., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.
Fischer, J., & Tognazzini, N. (2009). The truth about tracing. *Noûs, 43*(3), 531–556.
Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.
García-Carpintero, M. (2017). The philosophical significance of the de se. *Inquiry*, *60*(3), 253–276.
Gray, H., & Gray, G. (2007). Dimensions of Mind. *Science, 315*(5812), 619.
Godfrey-Smith, P. (1994). A modern history theory of functions. *Noûs, 28,* 344–362.
Goldman, A. I. (1986). *Epistemology and cognition*. Cambridge: Harvard University Press.
Haji, I. (1997). An epistemic dimension of blameworthiness. *Philosophy and Phenomenological Research, 57*(3), 523–544.
Kane, R. (1996). *The significance of free will*. Oxford University Press.
Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in medicine, 23*(1), 89–109.
Kripke, S. (1980). *Naming and necessity*. Harvard University Press.
Krishnan, A. (2009). *Killer robots: Legality and ethicality of autonomous weapons*. Routledge.
Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. Oxford University Press.
Levy, N. (2014). *Consciousness and moral responsibility*. Oxford University Press.
Lewis, D. (1979). Attitudes de dicto and de se. *Philosophical Review, 88*(4), 513–543.
Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing, 7*(1), 76–80.
Littlejohn, C. (2012). *Justification and the truth-connection*. Cambridge University Press.
Magidor, O. (2015). The myth of the de se. *Philosophical Perspectives, 29*(1), 249–283.
McKenna, M. (2012). *Conversation and responsibility*. Oxford University Press.
McLaughlin, P. (2001). *What functions explain: Functional explanation and self-reproducing systems*. Cambridge University Press.
Mele, A. (2010). Moral responsibility for actions: Epistemic and freedom conditions. *Philosophical Explorations, 13*(2), 101–111.
Mele, A. (2019). *Manipulated agents: A window to moral responsibility*. Oxford University Press.
Müller, V. (2020). Ethics of artificial intelligence and robotics. In: Zalta, E. N. (Ed.) *The stanford encyclopedia of philosophy* (Fall 2020 Edition). https://plato.stanford.edu/archives/fall2020/entries/ethics-ai/.
Nagel, T. (1974/2002). What is it like to be a bat? In: Chalmers, D. (Ed.) *Philosophy of mind: Classical and contemporary readings*. Oxford University Press.
Neander, K. (1991). Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science, 58*(2), 168–184.
Neander, K. (2017). *A mark of the mental. In defense of informational teleosemantics*. MIT Press.
Millikan, R. G. (1984). *Language*. MIT Press.
Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy, 86,* 281–297.
Neely, E. (2014). Machines and the moral community. *Philosophy & Technology*, *27*(1), 97–111.
Papineau, D. (1984). Representation and explanation. *Philosophy of Science*, *51*(4), 550–572.
Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford University Press.
Rudy-Hiller, F. (2018) The epistemic condition for moral responsibility. In: Zalta, E. N. (Ed.) *The stanford encyclopedia of philosophy* (Fall 2018 Edition). https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/.
Silva, P. (2017). The composite nature of epistemic justification. *Pacific Philosophical Quarterly, 98*(1), 25–48.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Fan, H., Sifre, L., Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature, 550*(7676), 354–359.
Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.
Sartorio, C. (2016). *Causality and free will*. Oxford Universty Press.
Sebastián, M. Á. (2012). Experiential awareness: Do you prefer it to me? *Philosophical Topics, 40*(2), 155–177.
Sebastián, M. A. (2018). Embodied appearance properties and subjectivity. *Adaptative Behaviour, 26*(5), 199–210.
Shepherd, J. (2014). The contours of control. *Philosophical Studies, 170*(3), 395–411.
Sher, G. (2009). *Who knew?: Responsibility without awareness*. Oxford University Press.
Smith, H. (1982). Culpable ignorance. *Philosophical Review, 92*(4), 543–571.
Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.
Stahl, B. (2006). (2006) Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology, 8,* 205–213.
Strawson, P. (1962/2003). Freedom and resentment. In: Watson, G. (Ed.) *Free will* (pp. 72–93). Oxford UniversIty Press.
Sullins, J. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6,* 23–30.
Trippi, R., & Turban, E. (1992). *Neural networks in finance and investing: Using artificial intelligence to improve real world performance*. McGraw-Hill.
Parthemore, J., & Withby, B. (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. *International Journal of Machine Consciousness, 6*(2), 141–161.
Peacocke, C. (2014). *The mirror of the world: Subjects, consciousness, and self-consciousness.* Oxford University Press.
Perry, J. (1979). The problem of the essential indexical. *Noûs, 13,* 3–21.
Pollock, J. L., & Cruz, J. (1999). *Contemporary theories of knowledge*. Lanham: Rowman & Littlefield.
Putnam, H. (1981). *Reason, truth, and history*. Cambridge University Press.
Recanati, F. (2007). *Perspectival thought: A plea for (moderate) relativism*. Oxford University Press.
Recanati, F. (2012). Immunity to error through misidentification: What it is and where it comes from. In S. Prosser & F. Recanati (Eds.), *Immunity to error thorugh misidentification: New essays*. Cambridge: Cambridge University Press.
Searle, J. (1990). Consciousness, explanatory inversion, and cognitive science . *Behavioral and Brain Science, 13*(1), 585–642. https://doi.org/10.1017/S0140525X00029587
Torrance, S. (2012) The centrality of machine consciousness to machine ethics. In: Gunkel, D. J., Bryson, J. J., & Torrance, S. (Eds.) *The machine question: AI, ethics and moral responsibility, AISB/IACAP World Congress 2012*.
Vargas, M. (2013). *Building better beings*. Oxford University Press.
Vikhar, P. A. (2016). Evolutionary algorithms: A critical review and its future prospects. In: *Proceedings of the 2016 international conference on global trends in signal processing, information computing and communication (ICGTSPICC), Jalgaon* (pp. 261–265).
Von Foerster, H. (1960). On self-organizing sustems and their ennvironments. In M. C. Yovits & S. Cameron (Eds.), *Self-organizing systems*. Pergamon Press.
Wallach, W., & Allen, C. (2010). *Moral machines teaching robots right from wrong*. Oxford University Press.
Watson, R. (1995). *Representational ideas: From Plato to Patricia Churchland*. Kluwer Academic Publishers.
Wedgwood, R. (2012). Justified inference. *Synthese, 189,* 273–295.
Zimmerman, M. (1997). Moral responsibility and ignorance. *Ethics, 107,* 410–426.