Philosophy of AI
Vol. 1, 2025, 41-58
10.18716/ojs/phai/2025.2276

AI, DECISIONS, AND THE REASONS TO BELIEVE: ETHICS-THROUGH-EPISTEMOLOGY APPROACH

Dina Babushkina

University of Twente, Faculty of Behavioural, Management and Social Sciences (BMS), Philosophy (WIJSB), Enschede, NL

Abstract The paper puts forward a notion of artificial intelligence (AI) as a *cognition technology* and centres on the role of AI systems as a sort of *epistemic plug* for the decision-making process by human *moral agents*. In this sense, the paper argues for *an ethics-through-epistemology approach to AI*. The question of responsibility is approached from the perspective of the decision maker and explains her motivational setup when deliberating about doing what is morally right. I start by arguing that understanding AI as a cognition technology forces us to re-conceptualize the moral responsibility for AI as the responsibility in the context of cognition. I next unwrap this claim and discuss three major elements in refocusing the philosophical discussion on the responsibility for AI: shifting focus (a) from actions to *decisions*, (b) from the question of imputability to the problem of *harm mitigation/prevention*, and (c) from ontological to *epistemic conditions*. Based on this, I argue that the responsible stance towards decision-making with AI presupposes *an obligation to evaluate reasons for actions*. I then analyse these in terms of *reasons to believe* in connection with the epistemic authority of AI as a cognition technology.

Keywords:
Psychology of responsible AI, responsible decision making, reasons to believe, ethics and epistemology of AI, cognition technology

1. Cognitive agency and AI

It is alarming how little we learn from the fact that artificial intelligence (AI) is *a technology of cognition*,[1] a technology the direct aim of which is to simulate and substitute different elements of human cognition. Mind-likeness is AI's overarching vision, the motivation that has been driving its development, and the main selling point. But claims about mind-likeness presuppose claims about cognition-likeness, i.e., likeness in the way human intellect relates to the world and the way artificial systems process data (analogous to the world of phenomena in human epistemology). I do not say these claims are true, I only say that they exist and persist. The ethical implications[2] of these claims are crucial, but still have not been researched analytically, especially with regard to claims about AI's (in)ability to substitute and supplement human cognition in making decisions.

The AI industry is encroaching on decision-making. AI systems already function as substitutes for various cognitive elements involved in this domain. However, decision-making

1 The way I use the term 'cognition technology' (or the 't echnology of cognition') different from the epistemic technology in, e.g., Alvarado's (2023) sense: while the former is used to grasp personal cognition aspect, the latter grasps the role of AI in scientific production.

2 I use "ethical implications" here in a broad sense, i.e., not merely limited to normative ethics, but also taking into account other areas of moral philosophy such as moral psychology.

is not a normatively neutral term. If decision-making means making decisions about actions that may affect the well-being of ourselves and others (which is true for AI), this is of prime interest for the ethical theory. Thus, normativity governing such hybrid decision-making processes should also be a central concern of theoretical philosophical ethics. We need to ask: What does such hybridization mean for responsible decision-making? If we consider that AI is a cognition technology that plugs into the decision process of humans, we are forced to rethink the way we approach moral responsibility in relation to AI and to conceptualize the responsibility for AI as a *moral* responsibility in the context of cognition (Baalen et al., 2021; Sebastián, 2021; Simon, 2015)[3]. Doing this theoretical work would allow us to identify new challenges and areas of inquiry. For instance, this paper invites us to assume the standpoint of moral psychology and to ask: How are we to conceptualize the motivational stance of an agent *who aims at the right choice* while deciding what to do and outsources a part of that decision-making process to an AI system? What consequences does such altering of the *decision-maker's epistemic profile have for the moral psychology* and especially for the theory of moral motivation?

This brings us to the intersection of ethics of AI that is motivated by the possibility of harm caused by AI-supplemented decisions and the epistemology of AI that governs the relationship between the outputs of AI systems and cognitive states (such as knowledge, belief, guess, or illusion) as well as the ways we arrive at these states (such as induction, deduction, intuition, or a leap of faith). Some of these are desirable, some are acceptable, and some are a bad way to interpret reality. The ethics of AI has to be informed by the epistemology of AI, which can give us a toolkit to evaluate such cognition technology as AI. What is missing now is a fundamental study, systematically analyzing and elaborating the groundings of the moral normativity about AI in AI's epistemic normativity. This is what I will refer to as *an ethics-through-epistemology approach to AI*.

This is not to say that ethicists have not been discussing AI's effects on decision, action, and responsibility. Those coming from theoretical ethics seem to be predominantly occupied with the fundamental questions of the conditions of moral agency (Behdadi & Munthe, 2020; Gogoshin, 2021; Hakli & Mäkelä, 2019) by trying to position artificial agents within that debate. The applied ethics often draws from the interdisciplinary sources and limits itself to more practical questions, specific applications of AI with the aim to create frameworks for policy and audit. But some insight can come from this limitation, as it becomes clear that the way the theoretical ethics discussion has been shaped is being challenged by the type of problems a decision-maker faces using AI.

Most notable lines of research, suggesting that the epistemology that have direct bearing on the ethics of AI, are on: opacity/transparency (some recent works: Von Eschenbach, 2021); reliability (Durán, 2021; Durán & Jongsma, 2021); explainable AI (Angelov et al., 2021; Dazeley et al., 2021; Durán, 2021; Fleisher, 2022; Mittelstadt et al., 2019; Páez, 2019; Zednik, 2021; Zednik & Boelsen, 2022); explainability with respect to fairness and injustice (Symons & Alvarado, 2022; Zafar et al., 2015; Zarsky, 2016); contestability (Ploug & Holm, 2022; Sand et al., 2021); responsibility, accountability & liability (Smith, 2021) (Kempeneer, 2021); and autonomy (Coeckelbergh, 2022). Researchers from various interdisciplinary subfields of ethics of technology are explicitly pointing to the link between some moral issues with AI and its epistemic limitations (Bjerring & Busch, 2021; Coeckelbergh, 2020; Grote & Berens, 2020; Simon, 2015).

A few scholars argue for a more systematic approach to the link between the two fields. For example Russo et al., (2023) argue for "epistemology-cum-ethics". This is an applied ethics framework for design and audit of AI systems that suggests including non-experts into the systems' assessment procedures to improve epistemic asymmetry, transparency, and understandability of AI's outcome for various stakeholders. Similarly, Mittelstadt et al., (2016) draw attention to cases where ethically problematic outcomes (unfair outcomes, transformative effects, non-traceability) of algorithms are a result of certain epistemic prob-

3 Interesting work is done within 4E cognition (Glackin et al., 2021) that aims at understanding significance for cognition of factors

lems (inconclusive evidence, inscrutable evidence, misguided evidence). They propose "…a prescriptive framework of types of issues arising from algorithms owing to three aspects of how algorithms operate [..,] as an organising structure based on how algorithms operate." This, again, suggests a distinct applied, audit-oriented stance.

It must be noted that my approach is not that of applied ethics. I suggest taking a step further towards a systematic account of AI that reveals and explains the inherent dependency of ethical normativity about AI on its epistemic normativity. Elements of this approach were developed in previous work (Babushkina & Votsis, 2022), where we claim that in order to understand how to ethically regulate AI, one needs to first understand how various AI systems generate their output and how those outputs relate to knowledge. This effectively subjects moral normativity, governing actions with AI, to the epistemic normativity governing AI's outputs, and we argue for the introduction of *epistemo-ethical constraints* (i.e., ethical constraints informed by the epistemic facts and norms) on decision-making with AI. The current paper develops these ideas further and takes a step towards *the conceptualization of AI as cognition technology*. The idea here is to look at what follows from the alteration of our cognition processes (as well as the introduction of the cognition elements that are not governed by familiar norms of human rationality) *for the motivational and deliberational aspects of the decision-making,*[4] *as the lived experience of a person*.

2. Responsibility for AI: From actions to decisions

To understand what AI supplementation of cognitive processes entails for motivation, we need to shift attention away from *actions and toward decisions*. While focusing on actions helps to reduce the unrealistic expectations from AI as an artificial agent and a possible object of reactive attitudes, it does not help us to understand how to position ourselves responsibly towards our decisions, assisted by AI systems. We need to focus on decisions—the internal states they involve and their rationality—if we are to make the discussion about responsibility for AI more relevant to decision-making rather than to action. The

external to the brain. Consider research on "hostile" cognitive environments and the way they shape an individual's cognitive agency. Glackin et al. (2021) discuss the concept of "cognitive scaffolding", i.e. an activity of arranging "our environment to facilitate our mental processes, by replacing the cognitive tasks and challenges that typically face us with more tractable ones" (Colombetti & Krueger, 2015; Saarinen, 2020; Sterelny, 2010). (Timms & Spurrett, 2023) point out that such scaffolding can be "hostile", i.e. "undermin[ing] or exploit[ing] the user while serving the interests of another agent" (Spurrett, 2024). Without going into the detail of the theoretical assumptions behind the 4E approach, this latter concept could be of interest for the study of AI as a cognition technology since, as such, AI relies on and facilitates the user's dependency on externalization of cognition and creates conditions for "deep scaffolding" and exposes the user to extreme cognitive manipulation.

4 A reviewer raised two important questions: (a) whether, according to my approach, the need to attend to the responsibility in decision-making arises specifically in the context of AI or whether—as follows from 4E cognition theory—AI is just one of the factors that shape our decision-making; (b) what are the limits of responsibility demands on evaluating every cognitive tool that influences our decisions (think, e.g. social scripts, habits etc.). My point is that AI is a cognition technology and has to be properly understood as such. The failure to do so comes with the risk of underestimating its invasive role in human decision. I do not deny that there are multiple factors that shape and constitute our decision-making processes, but AI is unique since it enters (and takes over parts of) our decision-making in previously unknown ways. The tangible epistemic risk associated with AI as a cognition technology is not only the loss of control over one's cognitive processes but also the loss of certain abilities and epistemic goods. These epistemic risks come with moral risk of undermining the epistemic conditions of the moral agency and thus preventing the agent from taking a responsible stance. This, however, does not mean that the agent is not responsible. In broad schema of things, there is nothing about the relationship between AI and human agency that constitutes an exception from responsibility (AI is certainly not the same as a gun at the head or a hurricane; the agent remains free and has the ability to get access to relevant information), and thus the agent herself may be blameworthy for the erosion of her epistemic state (except, of course, such cases as when the agent is indeed intentionally deceived or forced). As far as the connection between such factors as habits and social scripts, the moral self and moral responsibility goes, Bradley's ethical idealism (Bradley, 1962) has a lot to offer as it sees moral self as constantly evolving, also with respect to the beliefs, desires, and patterns of action. Bradley argues that the agent is blameworthy for bad habits (Babushkina, 2022) that influence her decision-making, since habits are a mere pattern of thoughts and actions which are, in principle, revisable. The same is appliable to social scripts (including, e.g., stereotypes) which—while are sometimes difficult to identify and could be invasive—are not deterministic since the agent has liberty to accept or reject a certain script.

difference is easy to overlook, but it is crucial. Bradley (1902, 1903, 1904), following Aristotle, Hobbes, and Hegel, brought attention to the fact that action is an agent's realised will or *a choice acted upon*. An action is an intention that becomes reality and changes the world according to the agent's will. A decision, on the other hand, *is not yet an action*[5], and so allows the agent to position herself towards the possible future state of affairs as well as her role in bringing them about. This is significant: the agent may evaluate alternatives and choose what will/will not happen or change her mind, e.g., in light of new evidence. The process of decision-making determines whether the agent will distance herself from something (e.g., because she judges this to be a wrong thing to do) and *not* bring that something about through her actions or whether she ties herself to something and makes it happen.

The difference between decision and action allows us to conceptualise the difference between what it means to be responsible for a decision (which may or may not result in an action) and what it means to be responsible for an action. Most discussion in the literature on the theoretical aspects of the ethics of AI has revolved around the latter; but recent developments in the applied field pave the road to a new direction, showing that, this limited focus makes it easy to underestimate, misunderstand, or even ignore the role, effects, and problems that AI-simulated cognitive elements play in the decision-making process, and often compels us to overestimate the role they should play in our actions. This does not mean that actions are not an important part of responsibility, but that we can no longer ignore the conditions of responsible decisions. Putting decisions in the spotlight allows refocusing the discussion about personal responsibility by shifting attention from 1) the problem of imputability of transgression/observance to *the mitigation of harm as a part of the motivational set up* of the decision-maker and 2) from ontological to *epistemic conditions* of responsibility, as they play out in her motivational set-up.

3. Responsibility set-up: From imputability to mitigation

To illustrate the discussion that follows, I will use a variation of a hypothetical example that I have introduced in another paper (Babushkina & Votsis, 2022). This example of decision-making relies heavily on output from an AI system that uses a deep neural network (DNN) to distinguish dogs (with a wolf-like appearance) from (actual) wolves. There are two important factors of this hypothetical model. First, it has a reported "high success rate". This is often claimed about AI in support of the algorithm's reliability and in order to boost potential decision-makers' trust in the outcome of the system. At the same time—and that is the second factor—the high success rate of the model is entirely contextual, i.e., it only applies to images exhibiting the same basic structure—something, we can safely assume, *the potential decision-maker is not aware of*. More specifically, were we to look closely at the algorithm's internal process of "feature extraction", we would find out that to classify an image as a wolf, the model relies on the image background, which contained snow in all images of wolves— consistently for the batch of images that was used at the pattern extraction phase.[6] This makes the model potentially unreliable and dangerous when applied to differently composed images.

This is the example that will be used throughout the paper [example Q]:

---

5 An objection may be raised against the distinction between decision and action, saying that decision is always a decision about an action. It is certainly true that a decision is a decision *about* an action (or omission as a form of action), but this points only to the propositional form: action is the propositional content of a decision as a certain metal state. However, if what is meant is that any decision inevitably leads to an action/omission, it gets a bit tricky. *Any* mental state (belief, desire, decision etc.) is followed by an action or omission by its subject, but that omission or action is not necessarily relevant to the content of that mental state. Think of situations when a subject (S) deliberating about A (action) and, say, decides to A, but is prevented from A-ing by circumstances independent of her will. In that case, the decision to A is followed by not A-ing which is not a result of S's decision to not A. Similarly, A maybe compelled by an external reason to B after she decided to A, but again, her B-ing is not connected to her decision to A. A decision is a state of mind and a choice between alternative actions; but a decision (or choice) is not itself an action.

6 For the explanation why the expression "pattern extraction phase" is more preferable than "learning" phase (Babushkina & Votsis, 2022).

> Imagine you are deciding what to do with an animal (An). If it is a dog, you will adopt it as a pet. You do not know what this animal is, and you cannot see it or an image of it. Instead, you are provided with a sort of decision plug, an AI system called BCSE ("Best Classification System Ever"), which could be advertised to you as a "reasoning assistant" or "decision-making assistant". The BCSE is given a photo of the animal (An) and processes it according to a DNN algorithm. As a result, it presents you with an output "An is 85% dog". This is the only information you have to decide whether you take An home to your family (remember, you do not know how (*) the BCSE arrived at its output).

Here is the question: How are you to responsibly position yourself to the choice before you? There are different ways you could do this. Were you to turn to philosophical literature on responsibility for advice, odds are you would discover that the mainstream discussion in the context of AI is heavily influenced by the traditional way philosophy problematises moral responsibility, i.e., with a heavy focus on the conditions for imputability[7]. The main question has always been: *under which conditions are we justified to hold someone morally responsible?* This involves identifying the negative conditions of responsibility, i.e., delineating cases when someone should not be held responsible. Notice that the focus of attention here is on the action and its moral properties—e.g., the action being wrong—and the moral subject. The question is whether we are warranted to ascribe the wrongdoing to that subject. I do not say "the agent" because we still need to find out if the subject carried out the action (i.e., whether she was the actual agent) so that we could establish the negative conditions of responsibility attribution. If the subject was indeed the agent, then we would want to determine the link between her will and the resulting harm. Clarifying these aspects is necessary in order to avoid such situations as wrongful accusation, scapegoating, or putting excessive demands on moral agency (supererogation). Identifying such negative conditions helps to formulate the norms of such moral practices as exculpation, vindication, or excuse.

The ethics of AI inherits this traditional focus on imputability from moral philosophy. As a result, in the interdisciplinary field of AI, we are witnessing a diverse discussion around such topics as[8]: Do artificial agents such as AI systems satisfy criteria of moral agency (Constantinescu et al., 2022)? *If not* (Coeckelbergh, 2020; De Cremer & Kasparov, 2022; Hakli & Mäkelä, 2019), then: What other reasons do we have to attribute responsibility to them and in what form (Tigard, 2021)? Does the fact that artificial agents do not meet the criteria of moral agency justify re-inventing the concept of moral responsibility in a way that would allow attributing it to them (Floridi & Sanders, 2004; Himmelreich & Köhler, 2022)? What does this moral unfitness of artificial agents entail for human moral practices? (For example: responsibility gap—an overview of the debate in (Nyholm, 2020; Santoni De Sio & Mecacci, 2021); more detailed discussion in (Da Silva, 2022; Königs, 2022); blame vacuum—in (Babushkina, 2020); retribution gap in (Danaher, 2016); diffusion of responsibility in (Bleher & Braun, 2022). How are we to distribute responsibility: Which elements of the complex action involving multiple parties, including programmers, direct users, professionals, legal entities) can be justifiably at-tributed to which parties (List, 2021; Neri et al., 2020)?

The accent on imputability sets the stage for the dialogue around responsibility for AI, bringing to the front a specific set of issues and expectations, while pushing the rest to the

7 I choose to talk about imputability in this context because it is a prerequisite for any attribution involved in accountability, liability, and blameworthiness. Especially with AI, the action-oriented discussion about responsibility is often concerned with the distribution of responsibility in the context of "many hands" problem (Coeckelbergh, 2020; Floridi, 2016; Strasser, 2022; Taddeo & Floridi, 2018). The point in this paper, however, is that we need to shift attention away fromfinding to whom we can attribute accountability/liability/blameworthiness, and to focus on the decision-making and its unique profile when it comes to moral responsibility. 8 There has been an explosion of literature on this topic in recent years. For a more systematic review of the debate on responsible AI (Behdadi & Munthe, 2020; Gogoshin, 2021; Hakli & Mäkelä, 2019; Loh, 2019).

side. The basic setting for responsibility discussion in *the imputability paradigm* is something like this:

> Responsibility is a post-action matter and is primarily motivated by the blame for a wrongdoing.

This has direct implications for the mental set-up of the moral agent who, before the action, is dwelling on her responsibility for using an AI system. The agent's mental set-up may look like this:

> W is morally wrong. Am I to blame for W? Is it fair that I am blamed for W?

Or often rather:

> Why am I not to blame for W? Are there any conditions that remove the burden of blame?

W can be an actual wrongdoing, which has occurred in real life, or a hypothetical wrongdoing, which has not yet happened but is being speculated, for example, for argument's sake: "Were W to happen, would I be the one to blame?" Or "Why would I not be the one to blame for what could have occurred in the future?" As long as the action is being *presented* as having occurred, we can discuss the actual or hypothetical attribution of wrongdoing. In either case, the basic mental set-up predisposes us to think about responsibility as something that occurs only after the fact. Not surprisingly, a frequent agent's concerns are about punishment[9] and how to avoid it.

In Q, what is crucial for the imputability setting is what happens after you have made up your mind and acted; more specifically, the potential harm caused by your bringing An to your family if An turns out to be a wolf. Someone in your family may get hurt, or the animal may be put to death. If that was all that mattered to you as a decision-maker, your motivational setup would look something like this:

> But what if An turns out to be a wolf, who is to blame for the damage? Do I feel that I can be justifiably blamed for someone getting hurt, given the conditions of my choice? Or, since I merely followed the recommendation of another agent (the BCSE), should the BCSE be dealing with the moral implications of the decision?

This is understandable, given that (1) you may not feel invested in the decision process enough to have control over it, and (2) because BCSE was advertised to you as an *epistemic authority*, capable of classifying things far better than a human agent.

The imputability setting for thinking about responsibility, even though important (especially from the legal point of view), is rather one-sided as it deals with *harm that is thought of as having occurred*. Contemplating your responsibility for possible harm is presented as you working through the scenario that evolved after the harm has occurred and has to be morally dealt with. If that is the only way to think about responsibility, this may give the false impression that the only thing we can do is to wait for the harm to happen and then distribute the blame. The traditional imputability setting fails to acknowledge another parameter of responsibility: harm which has not yet happened. You as a moral agent are not only positioned toward the harm that would have happened in the future if the animal you brought into your home turned out to be a wolf but are also positioned morally to the harm that may have been averted—i.e., you should not have adopted the animal, if

9 Hence the discussion about punishment and AI, including the need to punish artificial agents (Reichenbach et al., 2006); the (im)possibility thereof (Sparrow, 2007); comparison between the requirements for punishment between artificial and human agents (Guidi et al., 2021); and even the redundancy of punishment altogether.

that animal was a wolf. Plugging AI into the decision-making process makes this a crucial part because AI plugs *not* into our actions, but into our decisions, and at this stage, harm is not a given fact that has occurred, but a possibility that can—and should—be avoided/minimised *through the deliberation process*. The delegation of an element of cognition to an artificial tool simulating it, and thus, potentially weakening the agent's cognitive control over this part of decision-making, brings the pre-action harm mitigation into the forefront of responsibility discussion.

When AI is plugged into the deliberation process, it becomes an element in dealing with the hypothetical courses of action, including evaluating them under moral constraints such as avoiding harm and mitigating harm if it cannot be avoided ("Is this a wrong thing to do?" "Will acting this way cause harm?"). If this were the responsibility set-up in the example Q, preventing the morally wrong outcome would be of primary importance: for example, if An is a wolf, you have to ensure it is not adopted. Your motivation set-up in this case would be:

> How can I make sure that I will not bring a wolf to the house as a result of this decision-making situation?

Notice the crucial difference between the two responsibility set-ups:

> *The imputability set-up* centres around the agent's commitment to being called upon to answer to their actions and take the blame (under the imputability conditions) vs
>
> *The mitigation set-up* centres around the agent's commitment to making sure she is not doing what is wrong, i.e., around her commitment to prevent/mitigate harm.

In the imputability set-up, the agent's responsible stance equals to her readiness to, after the action, acknowledge something like this: "It was (morally) wrong for me to X; so, I am ready to take the blame/be punished". One possible development of the example Q is to admit: "I should not have brought An to my family; my children got hurt because of this, but since it was my decision, I should take the blame and live with the guilt" or "It was a bad choice to bring An to my home since this caused so much suffering to the animal, but since I am responsible for this, I will cover all relevant expenses and pay a penalty to the animal shelter". This reflects an expectation that, before acting, a responsible agent has a predisposition to accepting the blame for wrongdoing. This predisposition may be described like this: "I am ready to take the blame/be punished if my actions turn out to be morally wrong". But notice what is missing in the imputability set-up. This motivational story does not presuppose any commitment to preventing harm: the agent may risk making a morally wrong decision without doing much to ensure morally right outcome. In this way, the mitigation or prevention of possible harm is left to the chance of moral character of the person—whether she is willing to take *the moral risk*. This changes, however, if we shift attention from the action and the determinism that it presupposes (in the sense that acting predetermines a state of affairs) to the (yet) undetermined situation of making a choice. In the latter case, the goal is to minimise the moral risk. The responsibility in decision-making (which happens before the action) is a matter of active cognitive investment into the process of deliberation on the part of the decision-maker, to the extent that the agent has control over the outcome of this deliberation. This is why we deliberate as moral agents: to make sure that there are *no unnecessary moral risks*. The advantage of this point of view is the possibility to make *control over the process of deliberation about the future action* a part of a responsible stance towards using AI as a cognitive assistant, plugged into the deliberation process.

4. Conditions of responsibility: Ontological vs epistemic

The difference between imputability and mitigation set-ups amounts to the difference between the responsibility *in* decision-making and *for* decision-making. The responsibility *in* decision-making is a moral stance towards one's own cognitive (and non-cognitive) states that lead to a decision about a course of action. The responsibility *for* decision-making is a moral stance towards the result of such cognitive states after they have produced an action. It is not hard to see that responsibility in decision-making precedes and preconditions the responsibility for decisions and actions: in the world of rational agents, actions should be based on decisions. Shifting attention to responsibility in decision-making brings to light *the epistemic agency of the moral subject* and with it, the need to discuss *the epistemic conditions of moral responsibility* for decisions carried out with AI plugs[10].

Traditionally, the mainstream philosophical discourse about moral responsibility has centred around the ontological conditions. Philosophers have primarily asked: What *properties* should an agent have to qualify as an entity to which moral responsibility can be attributed? This includes, for example, being an appropriate object of reactive attitudes (Strawson, 2008); being able to give an account of her actions (in terms of reasons); and, in some cases, being punished. AI ethics has inherited this approach. As a result, addressing the possibility of imputability of wrongdoing to an artificial agent is normally seen as conditional on this agent having certain capabilities/properties that a non-artificial moral agent would have, such as sentience, empathy, and intentionality. The absence of such capacities makes them unfit for being the subjects of moral responsibility, effectively turning them into what Véliz called "moral zombies" (Véliz, 2021).

From this perspective, the main discussion in the Q case would centre around such topics as: (a) Did you lack the properties sufficient to constitute moral agency at the moment of decision? Did using the AI cognitive plug deprive you of these necessary properties? We could frame this discussion by finding out whether your choice was genuinely free or forced. (b) Can we attribute any of the properties sufficient for moral agency to the BCSE? For instance, is the artificial agent capable of acting with intent? How do machine intentional states relate to intentional actions?

Although the epistemic conditions of moral responsibility have been discussed in theoretical ethics, these discussions only recently gained more systematic attention (Campbell, 2019). But even so, debates about the cognitive states required for moral responsibility are mostly framed within the context of responsibility for the action, and the central question is: "What does the agent need to know/be aware of to justifiably be held accountable for their actions?" (Baum et al., 2022; Sebastián, 2021). Relevant to this is an awareness of the action and its consequences as well as an understanding of the moral qualities of the action. In the case of Q, if you were looking for potential ways to remove responsibility from yourself, you would need to demonstrate, for example, that you were not aware that bringing a wolf into the house could cause harm or that causing harm is morally wrong. Another strategy would be to argue that the purposeful design of your choice environment by third parties, where the AI plug was your only source of information about An, hid crucial information from you that was relevant to the future action. But even in this case, you would have to demonstrate that you have sufficient ground to trust the third parties or that you were forced to act. Either way, we are again looking at responsibility from the perceptive of imputability. What is at stake here are the epistemic requirements on imputability of an action, not of a decision. Of course, the abovementioned confirmations may also be seen to be relevant to the decision, but only as long as the decision was acted upon.

10 I am not saying that the ontological conditions are not relevant. I am only saying that when it comes to decisions aided by AI, we need to analyze and conceptualize the epistemic conditions in a sufficient manner. It is an interesting question how the ontological and epistemic conditions relate (but this is a matter for a separate research (Coeckelbergh, 2022), as it is not possible to deal with this question in proper manner in this paper.

However, what matters for deliberation, which is the core of the decision-making process, is a different question: *What does the agent need to know to make the right choice*?[11] Here, the focus is on the connection between the cognitive states of the moral subject/deliberator and the norm governing the rightness of the action. The motivational stance of the agent in this case could be spelled something like this:

> Given that I know I am responsible for the harm that may result from the actions following this decision (that means, among other things, that I will be called upon to answer for the harm and am ready to take the blame), what are the conditions under which I am able to make the right decision?

This involves understanding what parameters may prevent you from arriving at the right decision.

In the case of an AI-assisted reasoning[12] such as Q, the AI system constitutes a crucial condition for the decision and so is subject to the same check on compliance with the norm for the right action as any other key decision factor. In other words, BCSE plays a key role in what you will count as a reason to conclude whether adopting An is associated with certain risks of harm, and thus whether you evaluate a course of action as morally acceptable. What is at stake here is the connection between the wrongness of the future action and the cognitive states of the deliberator with regard to the deliberation process and the tools that she uses in this process.

Given this, your motivational setting in Q could be:

> Given that I know I am responsible for the harm that may result from the actions as a consequence of the decision I make with the assistance of BCSE, what are the conditions under which I am able to integrate the BCSE's outcome into my deliberation without precluding me from making the right decision?

As a responsible parent, you would be asking: "What do I need to know about An so that it is possible for me to make an informed decision whether it is suitable for adoption?" And given that the only thing you can know about An is that—according to the BCSE's output —An is "85% dog", your question should be:

> What do I need to know about the BCSE (and the way it produces statements about An) to be able to decide if An is suitable for adoption as a pet?

Such a motivational set-up entails an obligation to evaluate one's reasons for action. In application to AI-plugged decision-making, this constitutes an obligation *to evaluate the AI's output as such a reason*.

5. Obligation to evaluate reasons

When a moral agent makes a decision with an AI output as a parameter, she has an obligation *to evaluate the AI's output as a reason* for action [EYR]. This means that when you, as a moral agent, find yourself in the situation Q, you have an obligation to critically reflect whether the BCSE's output "An is 85% dog" is a good reason to adopt An. What sort of obligation is this?

Firstly, [EYR] is *a moral obligation*, and so prescribes what you should do to ensure a morally acceptable outcome of your deliberations. Secondly, [EYR] is *a rational requirement*,

11 In this respect, most relevant is the research on the human-in-the-loop (Monarch & Manning, 2021; Mosqueira-Rey et al., 2023) and meaningful human control.

12 For the discussion on what constitutes an assisting element in cognition (Clowes, 2019, 2020; Smart, Clowes, et al., 2017; Smart, Heersmink, et al., 2017).

which indicates what you ought to do *if* your deliberation is to be aligned with the principles of rationality. This implies that, if you do not evaluate your reasons, you are not being rational and are letting yourself arrive at a decision by other means. It could be faith, it could be emotion, or it could be the reliance on epistemic luck[13]. Furthermore, [EYR] is *an epistemic obligation*[14] because it relates to the grounds for believing. A part of this obligation is to take reasonable precautions not to be deceived. This implies making sure that the tool is not deceptive. Mitigating the risk of false belief in the epistemic authority of a decision support tool not only prevents such a belief from motivating one to act reprehensively, but it also helps to mitigate the risk of epistemic harm. Hence, the question: Since the BCSE gives an impression of epistemic authority to produce (undisputable) reasons for action, what is this authority based on? With this question, the agent is able to establish control over the deliberation process.

In its general form, [EYR] has this basic structure:

Obligation to B (beneficiary) about O (object) because of J (justification).

In the case of Q, the object of [EYR] is the ground for belief that the BCSE's output "85% dog" is a good reason for action. The obligation is justified by the vulnerability of those who are dependent on the outcome of your decision. The most interesting question here relates to the beneficiary: to whom do you have this obligation? This depends on your metaethical position concerning the nature of obligations, but several explanations are possible. First, based on the vulnerability estimate, anyone dependent on your decision is a potential beneficiary. Second, all moral patients are beneficiaries because of the universalizability principle (a Kantian solution). Third, you owe this obligation to yourself as a moral agent if we accept that all moral duties are essential duties to yourself (a Bradleyan solution, Bradley, 1962).

6. Reasons for action and reasons for belief

The obligation to evaluate one's reasons (including the AI's output) for action during the deliberation process is a necessary counterpart of accountability: if you are to account for your actions in a way that provides a morally satisfactory explanation, then the evaluation of reasons for what you thought was the right thing to do must have been part of your deliberation process.

The accountability requirement is an obligation to give an explanation for your actions. Different types of explanation are possible, but we can clearly delineate those that are not relevant. For example, we are not interested in a purely causal explanation, reconstructing the chain of events that led the agent to the action. This places explanation outside the agents' will (these still might be relevant to the explanatory story, for example, if we need to prove that the agent is not accountable). Our explanatory story needs to grasp the relationship between the action and the agent's internal states that led to the action.

When talking about the internal states, we are not interested in idiosyncratic explanations—let's call them *the sense-giving reasons*. These are a species of what is referred to as *motivating or explanatory reasons*[15]. They are motivating because they cite subjective facts that move the agent to action, and they are explanatory because they clarify why an action has happened rather than prescribe it. In contrast, justificatory reasons cite a norm that sanctions an action. Explanatory reasons usually cite an internal state of the agent, such as fear or excitement, which may not be apparent to the agent herself. Explanatory reasons are

13 For literature on epistemic luck (see, e.g., Pritchard, 2005, 2007, 2016); on the relationship between epistemic luck and moral luck (see, e.g., Pritchard, 2006; Statman, 1991)

14 On the concept of epistemic obligation/ duty (see, e.g., Feldman, 1988; Hall & Johnson, 1998).

15 E.g., (Smith, 2009).

sense-giving when they explain why the agent did what she did, *in her own eyes*; they create a sort of personal story about why the action was meaningful to the agent. In Q example, this could be a desire to please your child. Such reasons can be useful for understanding what the person thought she was doing, or how the action can be rationalized, but it does not necessarily explain how her internal states related to the *rightness of the action*.

Similarly to the explanation of action, the explanation of belief falls into two broader areas: the explanation why the cognitive subject believes that X, on the one hand, and the justification of her belief that X, on the other; and these two do not necessarily coincide. In the case of a moral choice—and to evaluate whether the subject's decision falls short of what is morally right—we want an explanation in terms of reasons that predisposed *the agent's belief that X was the right thing to do*. Therefore, our question is: "What grounds did you have to believe that it was right to F?" We can call these reasons—such that, by citing the agent's cognitive states, explain what she was doing to comply with the justificatory reasons for the action—*explanatory reasons for belief*.

Now, what we are interested in when we are holding someone morally accountable is the manner in which the agent's deliberation has contributed to bringing about the morally right action, that is to say, how in her case, the explanatory reasons to believe were related to the justificatory reasons for action. This means two things. First, moral accountability involves an obligation to explicate the link between one's cognitive states and the justificatory reasons for one's actions (thus linking the responsibility for actions and the responsibility for the decision). Second, the moral deliberator must first evaluate alternative courses of action from the perspective of the moral norms that justify them (we will return to this later).

In Q, we can represent the decision fork in the following simplified way:

> [Decision Fork # 1]
>
> If An is a dog, then I adopt An.
>
> If An is a wolf, then I do not adopt An.

The antecedent part of each judgement refers to a possible state of affairs, while the consequent refers to an action: both judgements condition an action upon a state of affairs. What is the nature of this conditioning? Is it causal? Hardly. An being a dog does not automatically lead to An being adopted. There is also no commitment to action: if An turns out to be a dog, there is no obligation to adopt it. You can change your mind or decide to act otherwise for another reason. Rather, we should understand this conditioning as a maxim or a subjective rule/principle of action. The maxim of Q is:

> I will adopt this animal *only if* it is a dog.

This maxim specifies conditions under which a certain action (in our case, the adoption of the animal) is acceptable for the agent. If we are to spell out the acceptability condition, it could look like this: "It is acceptable for me to adopt An only if it is a dog and not a wolf." Notice, however, that, because this is a maxim of a responsible agent, it specifies the acceptability in a universal manner and not in subjective, idiosyncratic terms; it is about conditions under which, for any moral agent in circumstances Q, it would be acceptable to adopt An. *This is about a normative reason for action*, i.e., the justification of the future action, but not its explanation. The norm that this reason expresses can be spelled out as "wolves should not be adopted", "dogs should be adopted", or perhaps more generally, "you should adopt this animal only if it is not a wild one". Appealing to such a norm is a way to justify adopting An, i.e., to explain why it is the right thing to do. However, the normative (or justificatory) reason does not give any information about how the norm "you should adopt this animal only if it is not a wild one" applies to your situation. It only says that something is the right thing to do under certain conditions, but it does not say what moved you to

act. There is a gap between motivating/explanatory and justificatory/normative reasons. It is possible that you, while knowing that it is (morally) right, chose to do what is wrong, or that you chose to do what is right but for the wrong reason. So, how can one make sure that one is doing what is right for the right reason, i.e., that the explanatory reasons are in line with the justificatory ones? One possible answer is: by analyzing the reasons that the agent has to believe that X (An is a dog). By choosing and evaluating such reasons, the agent is able to exercise control over the process of deliberation and make sure that the outcome complies with the normative rationality of her action.

The justificatory reason ("it is a dog and not a wild animal") specifies an ontological condition for your action (i.e., adopting An) meeting the criteria of (moral) rightness. Explanatory reasons, however, are tied to your belief that An is dog. The problem is that your belief may not correspond to the reality of what An is. So:

> a crucial part of the decision-making process in a case like Q, is to [try and] verify that An is a dog, i.e., that the deliberator's beliefs correspond to reality.

So, when answering "Why did you adopt An?", it would be insufficient to say: "Oh, I thought it is a dog". You would have to explain why you thought it was a dog. Think this way: what makes your choice of action in the Q case a success is the hard fact that An is a dog. You have only made the right choice to adopt An if it is, in fact, a dog. Now, how to do you get to this point of success? By having the correct belief about what An is. A part of this is making sure that you have a good reason to believe that An is a dog. You would need to supplement your belief that An is a dog with evidence to support this belief.

So, [Decision Fork # 1] is conditioning your actions on the truth-status of your belief that "An is a dog", so it should be revised as follows:

> [Decision Fork # 2]
>
> If I am right that An is a dog, then I have a good reason to adopt An.
>
> If I am wrong that An is a dog, then I do not have a good reason to adopt An.

This brings us to one of the most difficult issues in epistemology: How, in Q, can you verify that An is a dog? Since in Q, your only source of information about An is the BCSE, the question really is whether the AI's output gives a good reason to believe that An is a dog. Your decision situation now looks like this: if your deliberation leads you to adopt An, what makes this a success is An being a dog. This is an ontological condition, a hard fact that you cannot change. To bring your deliberation as close to this epistemic success as possible, you should—and, given the limitations of your situation in Q, also able to—do is evaluate BCSE's output "85% dog" (hereafter referred to as a *machine concept*) as a reason to believe (not only for you, but for anyone in your shoes) that An is indeed a dog.

Your decision fork should, therefore, be revised like this:

> [Decision Fork # 3]
>
> If the machine concept is a good reason to believe that An is a dog, then I have a good reason to adopt An.
>
> If not, then I do not have a sufficient reason to decide between the alternative choices.

In other words, *if the machine concept does not give you a good reason to believe that An is a dog, then you do not have enough information to make a decision about the action*. You need to further supplement your response to the question "Why did you do F?" by *explaining the epistemic authority of the BCSE*. This does not mean explaining why you believed it to be a good source of information (this would be an explanatory reason, merely describing the in-

ternal state by which you arrived at the conclusion, which could have been mistaken) but giving an epistemically justificatory reason, i.e., a normative reason that justifies the belief that the BCSE is a source of such knowledge.

7. Instead of conclusion: Cognition plugs and the relation to truth

So, does a machine concept of the type "X is Z" constitute a good reason to believe that Z is true? There are strong reasons to doubt that[16]. The "Black box" problem is well-known and widely discussed (Adadi & Berrada, 2018; Carabantes, 2020; Durán & Jongsma, 2021; Hamlyn, 1990; Von Eschenbach, 2021; Zednik, 2021). Understood as the opacity in how an algorithm produces its outcome, "black box" hinders critical evaluation of AI's recommendations. We cannot take AI recommendations uncritically, simply because we can never be certain that it refers the real-world phenomenon that we think it does.

Babushkina & Votsis (2022) develop the semiotic aspect of this by showing that human and machine concepts have radically different signification makeup and construct meaning by referring to different types of entities. As a result, even when they have the same linguistic expression (e.g., "dog") they in fact refer to different things: one to the real-world phenomenon and the other to a constructed and highly contextual property of a specific subset of digital information. This makes a user's inference from the machine output, which, regardless of its linguistic formulation, is a statistical expression of the type "85% dog; 15% wolf,"—to "therefore most likely dog" inherently problematic. It is even more problematic if we consider the difficulty with the interpretation of the concepts of uncertainty and probability that are used for the explanation of the way DNNs process data. A belief formed on the basis of an AI output (such as by a DNN) does not amount to knowledge because, when it is correct, this is largely due to epistemic luck. Relying on AI for information can actually harm your epistemic agency as it undermines your ability to acquire knowledge.

This does not mean that an AI system's output is not *a* reason to believe that X is true, but we should ask whether it is a *good* reason. Under certain conditions, even such actions as asking a random person on a street, crowdsourcing, or flipping a coin can yield some reason to believe that X is true, since all will give you some *information* about X that reflects various degrees of truth. But none of them is *epistemically reliable*—there is no guarantee that they will produce *knowledge*. Similarly, there is no guarantee that, when applied to a new case, the BCSE will produce the correct classification. Therefore, the outcome of using this method to make a judgment about reality is a matter of *epistemic chance*. This means that if you decide to go with it as a reason, then you are taking significant *epistemic risks*: despite the appearance of epistemic authority, this source of information may fail to bring you to a true belief.

What constitutes a success in a moral decision-making process (good choice) in a case like Q is whether it reflects the truth about what An is. Facts about An are an ontological given/constant, while your belief about An is an epistemic variable. As a responsible agent in control of the deliberation process, you have to make the epistemic variable A correspond with/adequately reflect what is ontologically given/constant, and not the other way around. The path to truth and facts lies through reasons to believe. Merely believing X to be true cannot be the normative reason to adopt An because if you are wrong, the reality will catch up with you. You cannot ignore reality; An is either a dog or not. There is no negotiation with reality, even if AI pushes us to think otherwise.

For us as cognizing agents, knowledge is a matter of relation to truth[17] and, with it, a matter of the alignment of our beliefs with how things are. The mind's relation to truth does not mean that we always get things right, but that if we get it wrong, there is an im-

16 There are numerous publications in computer science domain that discuss technical side of the problematic move from data to the conclusion via AI algorithms (Kenny & Keane, 2021; Menon & Williamson, 2018; Watkins et al., 2022; Zliobaite, 2015).

17 For an overview of theories of truth see (Glanzberg, 2021).

perative to alter what we think about the world and to revise and change the way we reached the wrong conclusion in the first place. That is to say, epistemic methods/tools are not intrinsically valuable; they are only valuable when they lead us to the truth. Some more traditional forms of automated information processing, such as syllogisms, guarantee arriving at a true conclusion from true premises. If your premises are true, then you can be certain that either your conclusions are true or you have made a mistake in the structure of your inferences. Other less categorical epistemic methods allow a degree of uncertainty, but we can normally spell out conditions under which they guarantee that we are not mistaken about the world. How does AI fit into our relationship with truth? Does AI have what it takes to guarantee that we will arrive at the representation of the world as it is? And, if we rely on AI to decide how to manipulate the world, does it have what it takes to guarantee that the result will be aligned with our epistemic goals of understanding the world?

The problem is that by itself, AI does not aim for truth. It is an information manipulation system with the basic goal of extracting a pre-determined subset from the available information. Thus, AI is not in the business of understanding how the world is; it is in the business of *fitting* information into pre-determined constraints. The basic principle of DNNs is determining how much similarity there is between different subsets of information. The criteria of the epistemic success of the machine in this case are not whether the output reflects reality, but whether the output reflects the fitting of the new information batch into the pattern derived from another information batch (Babushkina & Votsis, 2022). Truth-criterion only figures if superimposed by a human decision-maker while revising the machine output.

If we are to position ourselves responsibly towards the use of AI in decision-making, it should never be employed without additional truth-validating methods. In the case of Q, this means that you should either refrain from acting and seek alternative confirmation that An is a dog, or accept that you are taking certain epistemic chances and moral risks (since your deliberation may lead you to do what is wrong). If the goal is to minimize risks of harm, then we need more epistemic certainty than what the BCSE offers.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alvarado, R. (2023). AI as an epistemic technology. *Science and Engineering Ethics*, *29*(5), 32. https://doi.org/10.1007/s11948-023-00451-3

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5), e1424. https://doi.org/10.1002/widm.1424

Baalen, S., Boon, M., & Verhoef, P. (2021). From clinical decision support to clinical reasoning support systems. *Journal of Evaluation in Clinical Practice*, *27*(3), 520–528. https://doi.org/10.1111/jep.13541

Babushkina, D. (2020). Robots to blame? In M. Nørskov, J. Seibt, & O. S. Quick (Eds.), *Culturally Sustainable Social Robotics: Proceedings of Robophilosophy 2020, August 18-21, 2020, Aarhus University* (pp. 305–315). IOS Press.

Babushkina, D. (2022). The dispositional account of habits and explanation of moral action in F.H. Bradley. In J. Dunham & K. Romdenh-Romluc (Eds.), *Habit and the history of philosophy* (pp. 121–133). Routledge.

Babushkina, D., & Votsis, A. (2022). Epistemo-ethical constraints on AI-human decision making for diagnostic purposes. *Ethics and Information Technology*, *24*(2), 22. https://doi.org/10.1007/s10676-022-09629-y

Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, *35*(1), 12. https://doi.org/10.1007/s13347-022-00510-w

Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines*, *30*(2), 195–218. https://doi.org/10.1007/s11023-020-09525-8

Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology*, *34*(2), 349–371. https://doi.org/10.1007/s13347-019-00391-6

Bleher, H., & Braun, M. (2022). Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics*, *2*(4), 747–761. https://doi.org/10.1007/s43681-022-00135-x

Bradley, F. H. (1902). The definition of will. *Mind*, *11*(44), 437–469. JSTOR. http://www.jstor.org/stable/2248568

Bradley, F. H. (1903). The definition of will. *Mind*, *12*(46), 145–176. JSTOR. http://www.jstor.org/stable/2248175

Bradley, F. H. (1904). The definition of will. *Mind*, *13*(49), 1–37. JSTOR. http://www.jstor.org/stable/2248488

Bradley, F. H. (1962). *Ethical studies*. Oxford University Press.

Campbell, R. (2019). Moral epistemology. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Winter 2019 Edition). <https://plato.stanford.edu/archives/sum2024/entries/moral-epistemology/>

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI & Society*, *35*(2), 309–317. https://doi.org/10.1007/s00146-019-00888-w

Clowes, R. (2019). Immaterial engagement: human agency and the cognitive ecology of the internet. *Phenomenology and the Cognitive Sciences*, *18*(1), 259–279. https://doi.org/10.1007/s11097-018-9560-4

Clowes, R. (2020). The internet extended person: exoself or doppelganger? *Límite | Interdisciplinary Journal of Philosophy & Psychology*, *15:* 22. https://research.unl.pt/ws/portalfiles/portal/29762990/document_8_.pdf

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Coeckelbergh, M. (2022). *The political philosophy of AI: an introduction*. Polity.

Colombetti, G., & Krueger, J. (2015). Scaffoldings of the affective mind. *Philosophical Psychology*, *28*(8), 1157–1176. https://doi.org/10.1080/09515089.2014.976334

Constantinescu, M., Vică, C., Uszkai, R., & Voinea, C. (2022). Blame it on the AI? On the moral responsibility of artificial moral advisors. *Philosophy & Technology*, *35*(2), 35. https://doi.org/10.1007/s13347-022-00529-z

Da Silva, M. (2022). Autonomous artificial intelligence and liability: a comment on list. *Philosophy & Technology*, *35*(2), 44. https://doi.org/10.1007/s13347-022-00539-x

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299–309. https://doi.org/10.1007/s10676-016-9403-3

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, *299*, 103525. https://doi.org/10.1016/j.artint.2021.103525

De Cremer, D., & Kasparov, G. (2022). The ethical AI—paradox: why better technology needs more and not less human responsibility. *AI and Ethics*, *2*(1), 1–4. https://doi.org/10.1007/s43681-021-00075-y

Durán, J. (2021). Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artificial Intelligence*, *297*, 103498. https://doi.org/10.1016/j.artint.2021.103498

Durán, J., & Jongsma, K. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329. https://doi.org/10.1136/medethics-2020-106820

Feldman, R. (1988). Epistemic obligations. *Philosophical Perspectives*, *2*, 235. https://doi.org/10.2307/2214076

Fleisher, W. (2022). Understanding, idealization, and explainable AI. *Episteme*, *19*(4), 534–560. https://doi.org/10.1017/epi.2022.39

Floridi, L. (2016). Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Glackin, S. N., Roberts, T., & Krueger, J. (2021). Out of our heads: addiction and psychiatric externalism. *Behavioural Brain Research*, *398*, 112936. https://doi.org/10.1016/j.bbr.2020.112936

Glanzberg, M. (2021). Truth. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Summer 2021 Edition). <https://plato.stanford.edu/archives/sum2021/entries/truth/>

Gogoshin, D. (2021). Robot responsibility and moral community. *Frontiers in Robotics and AI*, *8*, 768092. https://doi.org/10.3389/frobt.2021.768092

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*(3), 205–211. https://doi.org/10.1136/medethics-2019-105586

Guidi, S., Marchigiani, E., Roncato, S., & Parlangeli, O. (2021). Human beings and robots: are there any differences in the attribution of punishments for the same crimes? *European Conference on Cognitive Ergonomics 2021*, 1–6. https://doi.org/10.1145/3452853.3452864

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist*, *102*(2), 259–275. https://doi.org/10.1093/monist/onz009

Hall, R. J., & Johnson, C. R. (1998). The epistemic duty to seek more evidence. *American Philosophical Quarterly*, *35(2)*, 129–139.

Hamlyn, D. W. (1990). *In and out of the black box: on the philosophy of cognition*. B. Blackwell.

Himmelreich, J., & Köhler, S. (2022). Responsible AI through conceptual engineering. *Philosophy & Technology*, *35*(3), 60. https://doi.org/10.1007/s13347-022-00542-2

Kempeneer, S. (2021). A big data state of mind: Epistemological challenges to accountability and transparency in data-driven regulation. *Government Information Quarterly*, *38*(3), 101578. https://doi.org/10.1016/j.giq.2021.101578

Kenny, E. M., & Keane, M. T. (2021). Explaining deep learning using examples: optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI. *Knowledge-Based Systems*, *233*, 107530. https://doi.org/10.1016/j.knosys.2021.107530

Königs, P. (2022). Artificial intelligence and responsibility gaps: what is the problem? *Ethics and Information Technology*, *24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

List, C. (2021). Group agency and artificial intelligence. *Philosophy & Technology*, *34*(4), 1213–1242. https://doi.org/10.1007/s13347-021-00454-7

Loh, J. (2019). Responsibility and robot ethics: a critical overview. *Philosophies*, *4*(4), 58. https://doi.org/10.3390/philosophies4040058

Menon, A. K., & Williamson, R. C. (2018). The cost of fairness in binary classification. *Roceedings of the 1st Conference on Fairness, Accountability and Transparency*, *PMLR 81*, 107–118. https://proceedings.mlr.press/v81/menon18a.html

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data & Society*, *3*(2), 205395171667967. https://doi.org/10.1177/2053951716679679

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. https://doi.org/10.1145/3287560.3287574

Monarch, R., & Manning, C. D. (2021). *Human-in-the-Loop machine learning: active learning and annotation for human-centered AI*.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Neri, E., Coppola, F., Miele, V., Bibbolino, C., & Grassi, R. (2020). Artificial intelligence: who is responsible for the diagnosis? *La Radiologia Medica*, *125*(6), 517–521. https://doi.org/10.1007/s11547-020-01135-9

Nyholm, S. (2020). *Humans and robots: ethics, agency, and anthropomorphism*. Rowman & Littlefield International.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w

Ploug, T., & Holm, S. (2022). Right to contest ai diagnostics: defining transparency and explainability requirements from a patient's perspective. In N. Lidströmer & H. Ashrafian (Eds.), *Artificial Intelligence in Medicine* (pp. 227–238). Springer International Publishing. https://doi.org/10.1007/978-3-030-64573-1_267

Pritchard, D. (2005). *Epistemic luck* (1st ed.). Oxford University PressOxford. https://doi.org/10.1093/019928038X.001.0001
Pritchard, D. (2006). Moral and epistemic luck. *Metaphilosophy*, *37*(1), 1–25. https://doi.org/10.1111/j.1467-9973.2006.00410.x
Pritchard, D. (2007). Anti-luck epistemology. *Synthese*, *158*(3), 277–297. https://doi.org/10.1007/s11229-006-9039-7
Pritchard, D. (2016). *Epistemology*. Palgrave Macmillan UK. https://doi.org/10.1007/978-1-137-52692-2
Reichenbach, J., Bartneck, C., & Carpenter, J. (2006). Well done, Robot! The importance of praise and presence in human-robot collaboration. *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 86–90. https://doi.org/10.1109/ROMAN.2006.314399
Russo, F., Schliesser, E., & Wagemans, J. (2023). Connecting ethics and epistemology of AI. *AI & Society*. https://doi.org/10.1007/s00146-022-01617-6
Saarinen, J. (2020). What can the concept of affective scaffolding do for us? *Philosophical Psychology*, *33*(6), 820–839. https://doi.org/10.1080/09515089.2020.1761542
Sand, M., Durán, J., & Jongsma, K. (2021). Responsibility beyond design: physicians' requirements for ethical medical AI. *Bioethics*, *36*(2), 162–169. https://doi.org/10.1111/bioe.12887
Santoni De Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: why they matter and how to address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x
Sebastián, M. Á. (2021). First-person representations and responsible agency in AI. *Synthese*, *199*(3–4), 7061–7079. https://doi.org/10.1007/s11229-021-03105-8
Simon, J. (2015). Distributed epistemic responsibility in a hyperconnected era. In L. Floridi (Ed.), *The Onlife Manifesto* (pp. 145–159). Springer International Publishing. https://doi.org/10.1007/978-3-319-04093-6_17
Smart, P., Clowes, R., & Heersmink, R. (2017). Mindsonline: the interface between web science, cognitive science and the philosophy of mind. *Foundations and Trends® in Web Science*, *6*(1–2), 1–232. https://doi.org/10.1561/1800000026
Smart, P., Heersmink, R., & Clowes, R. W. (2017). The cognitive ecology of the internet. In S. J. Cowley & F. Vallée-Tourangeau (Eds.), *Cognition Beyond the Brain* (pp. 251–282). Springer International Publishing. https://doi.org/10.1007/978-3-319-49115-8_13
Smith, H. (2021). Clinical AI: opacity, accountability, responsibility and liability. *AI & Society*, *36*(2), 535–545. https://doi.org/10.1007/s00146-020-01019-6
Smith, M. (2009). *The moral problem* (Reprint). Blackwell.
Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x
Spurrett, D. (2024). On hostile and oppressive affective technologies. *Topoi*. https://doi.org/10.1007/s11245-023-09962-x
Statman, D. (1991). Moral and epistemic luck. *Ratio*, *4*(2), 146–156. https://doi.org/10.1111/j.1467-9329.1991.tb00036.x
Sterelny, K. (2010). Minds: extended or scaffolded? *Phenomenology and the Cognitive Sciences*, *9*(4), 465–481. https://doi.org/10.1007/s11097-010-9174-y
Strasser, A. (2022). Distributed responsibility in human–machine interactions. *AI and Ethics*, *2*(3), 523–532. https://doi.org/10.1007/s43681-021-00109-5
Strawson, P. F. (2008). *Freedom and Resentment and other essays* (0 ed.). Routledge. https://doi.org/10.4324/9780203882566
Symons, J., & Alvarado, R. (2022). Epistemic injustice and data science technologies. *Synthese*, *200*(2), 87. https://doi.org/10.1007/s11229-022-03631-z
Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752. https://doi.org/10.1126/science.aat5991
Tigard, D. W. (2021). Artificial moral responsibility: how we can and cannot hold machines responsible. *Cambridge Quarterly of Healthcare Ethics*, *30*(3), 435–447. https://doi.org/10.1017/S0963180120000985

Timms, R., & Spurrett, D. (2023). Hostile scaffolding. *Philosophical Papers*, *52*(1), 53–82. https://doi.org/10.1080/05568641.2023.2231652

Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI & Society*, *36*(2), 487–497. https://doi.org/10.1007/s00146-021-01189-x

Von Eschenbach, W. J. (2021). Transparency and the black box problem: why we do not trust ai. *Philosophy & Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Watkins, E. A., McKenna, M., & Chen, J. (2022). *The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2202.09519

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). *Fairness constraints: mechanisms for fair classification* (Version 5). arXiv. https://doi.org/10.48550/ARXIV.1507.05259

Zarsky, T. (2016). The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118–132. https://doi.org/10.1177/0162243915605575

Zednik, C. (2021). Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & Technology*, *34*(2), 265–288. https://doi.org/10.1007/s13347-019-00382-7

Zednik, C., & Boelsen, H. (2022). Scientific exploration and explainable artificial intelligence. *Minds and Machines*, *32*(1), 219–239. https://doi.org/10.1007/s11023-021-09583-6

Zliobaite, I. (2015). *On the relation between accuracy and fairness in binary classification* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1505.05723