Philosophy & Technology (2024) 37:62
https://doi.org/10.1007/s13347-024-00752-w

RESEARCH ARTICLE

"Quasi-Metacognitive Machines: Why We Don't Need Morally Trustworthy AI and Communicating Reliability is Enough"

John Dorsch[1] · **Ophelia Deroy**[1]

Received: 9 January 2024 / Accepted: 25 April 2024

Abstract
Many policies and ethical guidelines recommend developing "trustworthy AI". We argue that developing morally trustworthy AI is not only unethical, as it promotes trust in an entity that cannot be trustworthy, but it is also unnecessary for optimal calibration. Instead, we show that reliability, exclusive of moral trust, entails the appropriate normative constraints that enable optimal calibration and mitigate the vulnerability that arises in high-stakes hybrid decision-making environments, without also demanding, as moral trust would, the anthropomorphization of AI and thus epistemically dubious behavior. The normative demands of reliability for interagential action are argued to be met by an analogue to procedural metacognitive competence (i.e., the ability to evaluate the quality of one's own informational states to regulate subsequent action). Drawing on recent empirical findings that suggest providing reliability scores (e.g., F1-scores) to human decision-makers improves calibration in the AI system, we argue that reliability scores provide a good index of competence and enable humans to determine how much they wish to rely on the system.

Keywords Ethics of AI · Trustworthy AI · Trustworthiness · Reliability · Science communication · Metacognition

✉ John Dorsch
johndorsch@gmail.com

Ophelia Deroy
ophelia.deroy@lmu.de

[1] Faculty of Philosophy, Philosophy of Science and Religious Studies, Ludwig Maximilian University, Munich, Germany

Published online: 03 May 2024

1 Introduction

Trust is an essential factor determining human decisions, one that shapes and constrains our actions towards individuals, organizations, and institutions. Crucially, this role may now be extending to certain sufficiently sophisticated artifacts. Since the emergence of automated systems (AS) and, more recently, systems of artificial intelligence (AI) as integral parts to the decision-making process, questions concerning trust in AS/AI have been at the forefront of research (for a review, see Glikson & Woolley, 2022).[1] This steadily increasing and encompassing role of intelligent machines has culminated in calls from both industry leaders and regulative bodies to develop so-called "trustworthy AI" (EU Guidelines, 2019). This paper aims to expose the wrongheadedness of this call and argues that reliability ought to be a more appropriate goal.

To understand the call for trustworthy AI, it is essential to recognize it as downstream from the earlier push for developing trustworthy automation (Sheridan & Hennessy, 1984), which led to the development of models describing the dynamics of human-machine interaction, models that remain deeply influential today (Lee & See, 2004). Moreover, careful analysis of the literature on trust in AI reveals its conceptual indebtedness to various models of *interpersonal trust*, such as those that result from Deutsch's (1977) groundbreaking research on conflict resolution. Such influences expose the underlying assumption that continues to shape contemporary discourse in AI ethics: trust not only mediates interpersonal relationships, but also relationships between people and automation. That said, it is unclear what it means to say an automated system is or should be trustworthy.

Through its conceptual analysis of trust, moral philosophy has yielded a relative consensus surrounding the proposal that trustworthiness is constituted chiefly by two components, reliability, on the one hand, and some additional X factor, on the other, with controversy over the exact complement to reliability (Simon, 2020). Thus, what might be called, the "trust equation" construes trustworthiness as essentially solving the sum of reliability plus X. But despite decades of debate hardly any agreement has emerged about the correct solution, with prominent views solving X in various ways by appealing to goodwill (Baier, 1986), a behavioral disposition to cooperate (Deutsch, 1977), an emergent property of social interaction (Luhmann, 1979), an emotional sensitivity to moral values (Lahno, 2001), or the encapsulation of the trustor's interests in those of the trustee (Hardin, 2002).

While it is unclear whether competing views can be reconciled into one vision of trustworthiness, we suggest that one common thread that strings them together is that moral trust, over and above mere reliance, involves a moral appraisal on the part of the trustor about the trustee. That is, moral trust involves either an affective evaluation or a cognitive judgement about the trustee as either morally agential, responsible,

[1] Though questions arise about how to delineate the terms "artificial intelligence", "machine learning algorithms", and "automated systems", they are treated as subsets of each other (AS - AI - ML). Until Section 3, we use the term "AI" to refer to any sufficiently sophisticated automated system that makes the human agent operating it vulnerable, so that questions of trustworthiness and reliability become a central concern to her use of the application.

accountable, or rational.[2] As such, an entity is morally trustworthy if this appraisal is correct relative to a context. For this reason, developing trustworthy AI is routinely regarded as a solution to the "the alignment problem" (Christian, 2020), namely, how to ensure that AS/AI, does not operate contrary to our values or facilitates actions that violate them.

That said, it is crucial to ensure that two undoubtedly orthogonal dimensions remain distinct regarding the assumption that trust mediates human relationships to automata. There is the descriptive dimension, how we *actually behave* toward AI, and there is the prescriptive dimension, how we *ought to* behave toward AI. If the goal is to describe and predict human behavior, then one might appeal to facts about socially enculturated agents who cannot help but employ trust when dealing with sufficiently sophisticated systems because trust is infused into the fabric of social life (Coeckelbergh, 2012). But if the goal is to develop ethical policies for the development and implementation of such systems, we ought to question whether this assumption is justified and ask whether moral trust ought to be employed in our relationship to machines or whether a related, but ultimately distinct concept, such as reliability (exclusive of moral trust), ought to modulate behavior toward automation.[3]

It is thus a matter of deep controversy whether trustworthy AI is part of the solution to the alignment problem or rather carves a larger cleft between technology and our values. While some researchers focus primarily on describing – and to limited degrees, justifying – our tendency to treat AI as trustworthy, often in an appeal to "diffuse forms of trust" that describe the blurring of lines between artificial agents and the humans responsible (e.g., Buechner & Tavani, 2011), some philosophers hold a positive stance toward trusting AI, offering minimal notions of moral agency and accountability, respectively (Floridi & Sanders, 2004). Meanwhile, other philosophers are critical of the call for trustworthy AI (Moor, 2006; Bryson, 2018; Ryan, 2020; Deroy, 2023), arguing this would result in normalizing anthropomorphism, a category mistake with disastrous consequences in the age of misinformation (O'Connor & Weatherall, 2019), wherein AI becomes deployed as a tool for societal, economic, and political manipulation.

For our criticism, we are solely concerned with autonomous systems, whose purpose is to facilitate human decision-making by taking autonomous action that assists the human in reaching her goal, which, in virtue of its automation, places the human

[2] It might be useful here to introduce common definitions for these crucial terms morally "agential", "accountable" and "responsible" (Floridi & Sanders, 2004). Moral agency is about being the source of moral action. Moral accountability is described by the capacity to change behavior to correct for moral wrongdoings. Moral responsibility, on the other hand, is about being the subject of moral evaluation and exhibiting the right intentional (mental) state. Finally, moral rationality is about being constrained by moral reasons (see Section 2 for more).

[3] As it concerns the issue of whether the distinction between trustworthy and reliable is a mere linguistic one, two forms of vulnerability remain nonetheless crucial for the practical philosophy of cooperation: one related to task-specific competence and another related to ethical norms (see Section 2). To argue for general applicability of this distinction, one might demonstrate (for example) that these are manifestations of the Kantian means/ends distinction, such that trustworthiness describes what agents ought to possess in the *as ends in themselves* mode, with reliability describing the *as means* mode (which, for humans at least, should be read as "never merely, but sometimes primarily" as means mode). Such an argument, if sound, would ground this distinction in fundamental principles of moral philosophy.

(and potentially other people) in a position of vulnerability, wherein concerns of trustworthiness and reliability become central to her use of the application. To that end, we argue that automated systems cannot be morally trustworthy. As it regards our proposal that AI does not need to be morally trustworthy because it can be designed to be reliable, we are exclusively concerned with machine learning applications that have been trained on a dataset about which there is a ground truth which is instrumental in establishing the reliability of the application.

Though reliability has been previously proposed as the correct target for ethical AI (Ryan, 2020), this paper aims to complement this research by conducting a normative analysis of reliability (Section 2), defending the claim that morally trustworthy AI is unnecessary (Section 3), and ultimately offering a framework for thinking about how reliability can mitigate vulnerability for humans in hybrid decision-making scenarios (Section 4). We argue that reliability is constituted by the consistency of competence, which justifies claims about the reliability of agents, biological or artificial, while demonstrating that consistent competence demands a capacity for procedural metacognition. We conclude with a summary and avenues for future research (**Conclusion**).

2 Reliability, Trustworthiness, and Vulnerability

In the technology ethics literature, there is no consensus around what reliability means. In developing his reliabilitst approach to epistemology, Goldman (1976) articulates reliability as a specific kind of tendency: "a cognitive mechanism or process is reliable if it not only produces true beliefs in actual situations, but would produce true beliefs […] in relevant counterfactual situations" (p. 771). Though Goldman's framing of reliability is epistemological, it can be modified to provide a condition for a reliable *agent or entity*: an agent/entity is reliable at accomplishing a task if she/it tends not only to be successful in the completion of the task in actual situations, but would also be successful in relevant counterfactual situations. This notion provides a definition of reliability that articulates its normative conditions as it entails correctness conditions: certain conditions must hold for an agent to be reliable, while a judgment about the reliability of an agent is only justified if such conditions hold.

This normative analysis of reliability discloses that reliability is a matter of consistent competence, competence in the past, into the future, and competence in relevant counterfactual situations. For example, Kawakami et al.'s (2022) qualitative study on the perceptions and experiences of users of AI-based decision support tools (for more on such tools, see Section 3 below) discusses a case wherein users perceive the tool as unreliable due to how its screening score appeared to change at random when the algorithm was run successively without any obvious change to its inputs. This is a clear case of reliability being undermined by inconsistency.

Notice that the fulfillment of these normative conditions establishes a prima facie reason for trusting an agent *as an agent*. So, before the relationship between reliability and trustworthiness can be articulated in detail, it will help to clarify what it means to say that AI can be an agent. Here, conditions on the part of agency are minimal, such that entities need only be capable of resolving "many-many problems" that

arise when acting (Wu, 2013). Agents are faced with having many inputs to act on and many ways to act on those inputs. Agency means having the capacity to resolve such problems by monitoring inputs and outputs and exercising control over which go with which. A typical action on the part of the AS/AI is making a recommendation or a prediction, which the human agent considers in her decision-making (see Section 3).

An agent's reliability thus serves as a *pro tanto* reason for trusting her, since knowing her to be reliable will justify, as far as it goes, making oneself vulnerable to her (for a discussion of the importance of vulnerability in trust-based relationships, see Baier, 1986). For example, if I know my cousin to be a reliable person, then I am justified in trusting her to perform some action, such as picking me up from the airport, unless, that is, I have a specific reason not to trust her. While reliability has some justificatory force in manners of trust, it is by no means sufficient, a point demonstrated by the trust equation: trustworthiness is reliability *plus* an additional X- factor. Even though I might know my cousin to be reliable, if I also have a reason to appraise her as immoral, her reliability can no longer serve as justification for the decision to be vulnerable to her. In fact, her reliability (as a stable character trait) might become a reason *not to trust her* as she might be relied upon to behave immorally.

To understand what justifies making oneself vulnerable to another agent, we need to place a richer gloss on the moral appraisal distinguishing moral trustworthiness from mere reliability. In line with much of the philosophical literature cited above, we understand that trusting entails an expectation on the part of the trustor that the trustee will do (or has done) the right thing for the right reasons relative to a context. Clearly then, an agent's reliability (either context-specific or as a general character trait), is insufficient to warrant trust, as this would demand knowing her behavior is constrained by "the right thing" and "the right reasons". For example, to believe that a data scientist is trustworthy is to believe that, in the context of data science, you expect she will do the right thing (e.g., ensuring the collected data is representative of the target population) for the right reasons (e.g., as this avoids bias), and to say that she is trustworthy is to say she does these things for these reasons in this context.

As it concerns AI, however, we are obviously not dealing with an agent that can do the right thing for the right reasons simply because, to the degree that AI can be said to have an intention[4], its intention cannot be grounded in the recognition of rational constraints, an appreciation of what is right versus what is wrong. Put differently, the intentions of AI cannot be grounded in *moral rationality*. Thus, the contribution of the present proposal is revisioning the rationality requirement typically demanded of artificial moral agency (e.g., Johnson, 2006) as applying to moral trustworthiness. This means, until such time as they can be designed to exhibit moral rationality, AI will never manifest moral trustworthiness.[5] Of course, this does not mean that vulnerability to these systems somehow disappears. On the contrary, the vulnerability is

[4] In the classical theory of action (Davidson, 1963), intentions are beliefs plus pro-attitudes. In the minimal account of agency articulated above, AS/AI can be said to have intentions if they derive solutions to many-many problems (minimal beliefs) and have programming for solving such problems or similar problems (minimal pro-attitudes).

[5] It is important to note the possibility of restricting trust solely to epistemic matters, wherein an argument could be made it is potentially justified to trust AS/AI *solely as an epistemic source or epistemic tool* (see

often exacerbated specifically because these systems are incapable of forming intentions that conform to normative knowledge.

If AI cannot manifest trustworthiness, then we ought to pursue all appropriate means to mitigate the vulnerability incurred by human-machine interaction. This is because vulnerability to AI can never be justified since its actions only ever support a *pro tanto* reason to trust (namely, reliability), which is inevitably defeated by knowledge that AI's intentions cannot be grounded in a recognition of rational constraints, so that vulnerability is a necessary consequence of hybrid decision-making situations. Below we discuss this problem in more detail (particularly, how to mitigate it by developing machine learning applications to meet the normative conditions of reliability), but for now, we need to expose why designing systems to manifest moral trustworthiness would not actually mitigate the vulnerability that arises in high-stakes decision-making environments.

3 Why Morally Trustworthy AI is Unnecessary

For the purposes of our argument for why morally trustworthy AI is unnecessary, we are exclusively concerned with machine learning applications.[6] In particular, we are focused on those applications deployed in medium to high-stakes environments, exemplified by the real-world case of decision support tools in child welfare services to assist child maltreatment screening decisions (Kawakami et al., 2022), specifically the Allegheny Family Screening Tool (AFST) (Vaithianathan et al., 2017). Such applications promise to bring more balanced results to this decision-making process by mitigating biases in human decision-making (Lebreton et al., 2019). But since machine learning algorithms are susceptible to their own biases, ones quite distinct from those of humans (De-Arteaga et al., 2020), a growing body of research has called for human-AI partnerships (Kamar, 2016). The core program is that humans and AI ought to build upon each other's strengths and compensate for their weaknesses (Bansal et al., 2021; cf. Green & Chen, 2019). Such hybrid environments are the target of our analysis which exposes how moral trustworthiness is unnecessary and reliability is sufficient.

AFST is a predictive risk modeling tool that generates a screening score for incoming calls of child mistreatment allegations by integrating and analyzing hundreds of data elements. The score aims to predict the long-term likelihood that child welfare services will be involved in a particular case. Should the score meet a specific threshold, an investigation is mandatory. But in all other cases, the decision is left to the

Alvarado, 2023). But here we are interested in trust in general, which, as argued above, is an implicitly moral notion. Henceforth, trust is always meant as moral trust.

[6] We are no longer concerned with all AS, since non-machine-learning AS (lacking, e.g., processes for updating internal models and minimizing error), cannot meet the normative requirements for reliability established in Section 2. Henceforth, by "AI" we mean autonomous machine learning systems for decision support. As our argument concerns AS in general, it follows from our view that we should design any AS that raises similar vulnerability issues to realize quasi-metacognition and to communicate its reliability to the relevant human agents because this would be sufficient to mitigate the vulnerability incurred by its deployment (see Section 4).

human agent, "the screener", and the score serves as an additional piece of information to assist the decision of whether to open an investigation (Samant et al., 2021), thus making clear the vulnerability caused by such environments. Since an inaccurate score has the potential to affect the screener's decision on whether to begin an investigation, the screener, the affected children, and their caretakers, are vulnerable to the AI's contribution to the decision space. This is not to say that the worry is that the screeners over-rely on the tool. Given how screeners consider details that AFST does not have access to, such as the potential motives of callers, cultural misunderstandings, and so-called "retaliation reports", wherein parents in custody battles repeatedly report one another during a dispute (Kawakami et al., 2022), studies have consistently shown that overreliance is not an issue. The point is that screeners must *calibrate their reliance* on the AI's recommendations that are not always accurate (more on this calibration process below).

Thus, the people involved are in a position of severe vulnerability that can be diminished by designing AFST to be as reliable as possible in analyzing, integrating, and summarizing the relevant data elements. One might wonder, however, whether reliability is sufficient to deal with this problem of vulnerability or whether AFST ought to be designed to be trustworthy. At first glance, this certainly seems like a noteworthy goal, since trustworthiness is a categorical reason to trust, one which would justify this vulnerability. But, given what was discussed above, what exactly does it mean to say that a machine is trustworthy? Recall that the consensus in the philosophical literature is that trust entails a moral appraisal on the part of the trustor about the trustee and trustworthiness entails this appraisal being correct relative to some context. Though attempts to instill a sense of morality into AFST would likely be in vain, it might still be designed to at least *simulate* degrees of moral sensitivity, and doing so could be in line with developing trustworthy AI (i.e., so long as simulation is part of the agenda).

How is that possible considering the above discussion about the nature of trust? We can list at least four possibilities: an AI might be designed to simulate (a) moral agency, (b) moral responsibility, (c) moral accountability, or (d) moral rationality – either of which, it might be argued, would lend support to the proposal that the AI is morally trustworthy. Concerning (a) moral agency, if this is understood in a minimal sense, that of being the source of moral consequences (Floridi & Sanders, 2004), then systems like AFST are already moral agents, since they produce the screening score, which, once used, has moral consequences. Thus construed, moral agency is better thought of as a *precondition* on the problem of vulnerability introduced by use, rather than one of its solutions. Of course, more robust notions of moral agency exist, wherein, for instance, the agent must act on behalf of moral values (Cherkowski et al., 2015), but in the framework deployed here, the capacity to act on behalf of moral values is a criterion for moral responsibility.

Another avenue for making AI trustworthy would be to develop it to simulate (b) moral responsibility, which would entail, given how responsibility is distinguished from accountability, that human agents are justified in evaluating AI behavior as either morally blameworthy or morally praiseworthy (see Hieronymi, 2004 for this distinction as it applies generally). Note, however, that until AI can be designed to exhibit an awareness of right and wrong, it would be a conceptual mistake to transfer

our praise or blame onto an artificial agent. The argument here is not that our practices of blame and praise are biased or flawed (see, e.g., Longin et al., 2023 and Porsdam et al., 2023, for empirical and normative discussion of this point). The argument is that our practices demand that the goal of moral evaluation is to shape the recipient's conscience, structure her cognitive repertoire to better align with shared values. Until AI possesses an emotional and social self, AI cannot be an appropriate target of such rich sociocultural mind-shaping practices (for research into such practices, see Zawidzki, 2013 in philosophy and Heyes et al., 2020 in cognitive science). Setting aside whether it is feasible to build an AI that can possess such a complex self, the question arises whether having such a self would indeed contribute to mitigating the problem of vulnerability that arises in high-stakes decision-making situations. This issue will be dealt with below, after we have made sense of how an emotional self is entwined with moral rationality (see Section 3.4).

Thirdly, we might design AI to simulate (c) moral accountability, which would entail it possess the capacity to correct its behavior to comply with moral norms. In the case of AFST, the application might be designed to update its predictions due to mistakes of a moral kind, which might be controlled by the screener. For example, one common issue that screeners see with how AFST's produces its score is its tendency to increase the severity of the case as more allegations are made. In the event of escalating retaliation reports, this will have a pronounced negative effect on the AI's performance. If the system were to learn not to increase severity due to retaliation calls, then, proponents of trustworthy AI would claim, the AI would be simulating a minimal form of moral accountability.

If this is all there is to moral accountability, then, like moral agency above, moral accountability is already implemented in AFST: as a statistical model, it employs representations of its predictive error and utilizes them to update its predictions. But despite being so designed, it is hard to make sense of how this autocorrection process, ubiquitous in machine-learning applications, makes AFST trustworthy in a manner distinct from how it simply makes AFST more reliable. This process only ensures AFST is a consistently competent agent (see above, Section 2).

Consequently, two takeaways emerge. First, designing AIs to exhibit a minimal notion of moral accountability only makes AIs more trustworthy in virtue of how it makes them more reliable. If this is correct, questions arise about why the scientific and political community ought to call AIs trustworthy if reliability is a sufficient descriptor, especially if this might bias people into believing moral appraisals of AI are justified. To prevent such epistemically dubious behavior, we wish to introduce a principle to guide communication:

> The Principle of the Non-equivalence of Trustworthiness and Reliability (PNTR)

To avoid unnecessary moral entailments or implications that come with speaking about trust, one ought to speak of trustworthiness (in manners of communication) only when trustworthiness means something over and above reliability.

PNTR say that if the notions "trust", "trustworthy", and "trustworthiness" can be replaced without semantic loss by the notions "rely", "reliable", and "reliability"

respectively, then one ought only to employ the latter for the purposes of scientific and policy communication.[7]

Second, since notions of minimal moral accountability only make AI more trustworthy in virtue of how it makes it more reliable, it is too thin to be of any use to make sense of how AI might mitigate vulnerability in a distinctly trust-involving manner. To examine whether this would be possible, we need to turn to a more robust notion of moral accountability. By "distinctly trust-involving", we mean to say the AI would be capable of simulating moral trustworthiness, as it is understood above as simulating (d) moral rationality (i.e., doing the right thing for the right reasons), so that talk of such forms of trustworthy AI would not fall under the jurisdiction of the above principle. The question then becomes whether designing AI to exhibit moral rationality would assist in mitigating the vulnerability that arises in high-stakes hybrid decision-making environments. But before this issue can be addressed, we must consider underlying details concerning what it would mean for an agent to possess moral rationality.

3.1 The Cognitive Conditions for Moral Rationality

Philosophers have long emphasized that capacities for rule acquisition and compliance, even rules that ought to be interpreted as moral, are insufficient for moral rationality as this requires a sensitivity to moral *norms*, which are irreducible to rules (for a review, see Haugeland (1990); and for the relationship between accountability, normativity, and rationality see Strawson, 1962). By drawing clear lines between mere rule-following and the sophisticated norm-following behavior, Gibbard (1990) develops a naturalistic theory on the origin of normativity, proposing two distinct cognitive systems that enable (i.e., are sufficient for) normative behavior to manifest: (i) an emotional system and (ii) a conceptual system.

To illustrate what distinguishes the two systems, Gibbard introduces the elaborate ritual that ensues whenever two dogs meet on neutral ground. Their interaction follows regular patterns that Gibbard believes exhibits a form of rationale that emerged through the pressure of natural selection for coordinating behavior among conspecifics: "…in this special sense, the beasts 'follow [norms]' for social interaction. They have not, of course, decided to conduct themselves by these [norms]…" (p. 69). For (ii) the conceptual system, which underpins the capacity to decide to conduct oneself according to a norm, requires a capacity to represent a norm as such, which, according to Gibbard, depends upon language and propositional thought, which produces mental representations *as* representations (i.e., representations whose content is determined by correctness conditions, e.g., Peacocke, 1992). Essentially, conceptual capacities, according to Gibbard, have the effect of enriching an agent's decision-making repertoire, allowing for motivation based on norms *as such*.

[7] Thus, we respond to the "so-what objection" (Mainz, 2023) by rejecting the claim that it is trivial to point out that AI cannot be trustworthy. Simply put, *what we say matters*: calling AI trustworthy, when it is not, propagates misinformation and engenders miscommunication that could result in mistrust in science (National Academies of Sciences, 2017).

While conceptual capacities distinguish the normative behavior of humans from the normative behavior of non-human animals, it is (i) the emotional system that distinguishes the behavior of non-human and human animals from that of non-normative, mere rule-following entities. Gibbard describes the emotional system as enabling the agent to "be in the grip of norms", a process of internalization, constituted as much by a specific behavioral profile as it is by a species of emotional receptivity: "a norm prescribes a pattern of behavior, and to internalize a norm is to have a motivational tendency of a particular kind to act on that pattern" (p. 70). For Gibbard, tendencies are evolutionary adaptations for coordination, while emotion (or affect, more generally) explains the motivation to perform the coordinated act: "When a person's emotions tend to follow a pattern in this way, we can say that the person internalizes the norm that prescribes the pattern" (p. 71).

Consequently, Gibbard draws two sharp lines, one between normative and non-normative entities and another between mere normative entities and human agents. To be a normative entity, one must be capable of internalizing a rule, thereby transforming the rule into a norm, which requires possessing either a motivational tendency to enact specific behavioral patterns that evolved for social coordination or the conceptual capacity to represent norms as such. In other words, normative agency is about having the right kind of intentional state driving the rule-following behavior, namely either (i) an affective state of attunement to the sociocultural environment or (ii) a belief state constituted by the correctness conditions describing the normative behavior. Thus, Gibbard's theory of normativity is particularly useful for present purposes because it helps articulate boundaries around moral rationality. To be capable of exhibiting moral rationality, one must be sensitive to moral norms, a receptivity which is irreducible to a capacity for following rules: for rules to qualify as norms, they must be either *felt* or *understood*.

3.2 AI Is Not Morally Rational

Applying Gibbard's naturalistic theory on the origin of normativity, we can safely say human behavior is guided by norms in a manner distinct from how AI might be said to be guided by rules. Our responsiveness to norms arises from *conceptual capacities* that enable the representation of norms as such, sometimes referred to as capacities for "representing as" (Dretske, 1988; Fodor, 2008). It is widely believed that these capacities hold a special status as exclusively human, making possible a form of metacognition known as "metarepresentation" (Proust, 2013), that is, the capacity to represent representations as representations, as mental entities with correctness conditions that can misrepresent, and thus represent beliefs as beliefs, which could be false; perceptions as perceptions, which could be inaccurate or illusory; and norms as norms, which could be inapt (Evans, 1982; Peacocke, 1983; McDowell, 1994). Importantly, the capacity to metarepresent is much more sophisticated than the capacity to explicitly represent (for representing rules explicitly in machines, see Sharkey, 2020), since metarepresentation is often argued to enable *reflective self-knowledge* (see Heyes et al., 2020 for an account of this). To be clear, the day may come when AI acquires the capacity to reflect and acquire self-knowledge, but the AI of today is a far cry from such sophistication, and so describing AI as capable of

accepting norms, and thereby exhibiting moral rationality in this sense, is clearly a conceptual mistake.

Analogous to the issue above of claiming AI has metarepresentational capacities, the claim that the AI of today can undergo emotion, such that it would be warranted to say it is *motivated* to comply with moral norms (thus exhibiting morally rational in this sense) would be a conceptual error. Despite the progress that affective psychology has made in recent years, for example, with the explanatory power of the neo-Jamesian interoception-based theory of emotion (Tsakiris & De Preester, 2018), an exhaustive account of the antecedents of emotion remains to be discovered. That said, a relevant consensus surrounds the *multiple components view* of emotion, wherein emotion is composed of various essential parts, one of which being the physiological component (Scarantino & de Sousa, 2018). Thus, we can say (without invoking feelings) that emotion demands an internal milieu, whose regulation is essential for the agent's survival, requiring a delicate balancing of metabolic variables – or, simply put, molecular turnover – something which AI does not do.[8]

As a result, we have convincing reasons to consider thinking of AIs as accountable in a trust-involving sense as a conceptual mistake. Though AI can be designed to monitor and adjust predictive weights to comply with moral rules, rule compliance is not norm compliance. Put simply: a fundamental difference exists between Jupiter obeying the laws of physics as a *consequence of natural law* and the young child choosing not to steal *for the reason* that it is morally wrong, the former of which is closer to artificial systems updating their models *as a consequence of operational rules*. Of course, in between these extremes is the norm-guided behavior of animals that enact behavioral tendencies *in virtue of being motivated* by emotional sensitivity to regulatory patterns, something which, of course, AIs cannot do.

3.3 Even Morally Rational Artificial Agents Do Not Help

The specific issue that our discussion aims to shed light on is how it is deeply unclear whether designing AI to be morally rational, such that they possess either (i) the emotional system or (ii) the conceptual system and so would be deserving of the appellation "trustworthy AI", would mitigate vulnerability arising in high-stakes hybrid decision-making environments. Below we begin with (ii) and then move on to (i), arguing that neither system implemented in AI will mitigate the relevant vulnerability.

To begin exposing this issue, we first need to make sense of what this vulnerability entails. The relevant vulnerability arises because of the need, on the part of the human agent, to calibrate her reliance on the AI's performance (see Lee and See (2004), who refer to this as "trust calibration"). Put simply: optimal reliance calibration means the human relies on the AI when its advice is correct (or more correct if the outcome

[8] Notice that an appeal to metabolic processes is not necessarily opposed to a functionalist account of emotion. A functionalist could accommodate this demand by appealing to functions, such that properties of the environment serve as input, while properties of the boundaries that constitute the agent serve as output. One would then need to provide a convincing case for how these processes are constitutive of the AI's basic interaction with the environment. Of course, it is unclear whether this would be sufficient for claiming the agent's emotions are constituted by feelings, as many theorists would demand, but, at the very least, it could theoretically implement physiological functions.

is not categorical) and deviates from the AI's advice whenever incorrect (or more incorrect). Were this optimum realized, the relevant vulnerability would disappear. Obviously then, solutions to this problem of vulnerability ought to aim at facilitating optimal calibration of reliance, such that the screener can more easily distinguish correct from incorrect cases.

Starting with (ii) the conceptual control system and sticking with AFST, we might ask whether the AI's capacity to represent norms *as such* would facilitate the screener in determining whether the AI's score is an accurate representation of the actual likelihood that intervention is needed. Though it is difficult to imagine what such a system would look like, we might provide a simple answer by pivoting to how humans are able to represent norms as such, and so, for the sake of the argument, we might envision simply replacing AFST with a human. However, replacing the AI with a full-blooded moral agent, even a perfect one, will not necessarily facilitate the screener's decision about whether the score should be relied upon or deviated from.

This is because, of course, having a capacity to regulate your behavior around moral norms as norms does not assist another agent in determining whether you are providing accurate recommendations. It does not even entail an increase in reliability, since morality-based regulatory capacities do not offer any advice on how to analyze and interpret the factors that determine whether children are at risk of abuse, and so this capacity does not facilitate the screener's calibration. Rather, what is needed for this task is detailed knowledge about the relevant causal antecedents predictive of child maltreatment and neglect. Indeed, we can envision an amoral agent that is proficient at determining the relevant causal factors and assigning appropriate weights based on statistical regularities, and hence why AI is usefully deployed in these environments.

Because the AI would be capable of representing norms as such, we can envision AFST as providing information about the moral norm that it adhered to in producing the screening score (e.g., "Good parents ought not to have criminal records"). While this might be helpful in determining whether the score should be relied upon, and so might assist in mitigating the vulnerability for all stakeholders, the AI can be designed to yield this information without thereby having the capacity to base its decision on moral norms as such. That is, the AI might be able to offer this "reason" without a capacity for metarepresentation. It could simply be designed to highlight the information relevant for its determination or, somewhat more sophisticated, be designed to generate a written summary of the relevant factors influencing its decision, a summary that might be fine-tuned to appear like a reason (or might be fine-tuned to be a reason to the right kind of human agent).

Moving on to (ii) the emotional system, which enables agents to be motivated by norms, we might likewise ask whether designing an AI to be morally rational in this manner would help alleviate the problem of vulnerability in high-stakes hybrid decision-making environments. As above, the question is whether this would lead to optimal calibration, that is, facilitate the screener's ability to discriminate accurate from inaccurate screening scores. But, once again, it is unclear how equipping AI, such as AFST, with emotional sensitivity will facilitate determining whether the screening score should be relied upon.

For example, if AFST were designed to report the emotion that motivated its screening score, the screener would be tasked with speculating about the explanation behind the emotion. If the screener knew it was fear, say, rather than sadness, that guided AFST to score a particular case as 15 out of 20, it would be hard to make sense of how exactly this is useful other than as a prompt to initiate some detective work about why fear was influential. Rather, the AI might be designed to provide an explanation, rather than an emotional abstraction, which, as above, the AI could be designed to do, without undergoing or even reporting an emotion.

3.4 Where Does This Leave Us? The Significance of Trustworthiness and Why AI Still Does Not Need To Be Morally Trustworthy

Before discussing conceptual details around the distinct societal role of trustworthiness compared to reliability, let us illustrate this key notion with a quotidian example. Imagine asking your partner to give you a haircut even though they are not professionally trained and have never cut hair before. Obviously, it makes little sense to say they are reliable at giving haircuts, since they have never attempted to cut hair. Now, let us say your partner is morally trustworthy. What does this mean? It means, though they are not reliable at cutting hair, they are nonetheless reliable at conforming their behavior to certain moral norms, either by way of explicitly representing those norms via conceptual capacities or by way of being motivated to adhere to those norms via emotional capacities (and usually by way of both).[9] For instance, they will cut your hair carefully for the reason that they respect how you would feel if they acted carelessly. In these cases, we can talk about trustworthiness that is irreducible to reliability, since it makes little sense to say the trustee is reliable at performing the task, but it does make sense to say that the trustee's character – her trustworthiness – justifies the trustor's decision to have her perform the task.[10] Put differently, trustworthiness is society's solution to the problem of how to bridge the gap in task-specific reliability, and it tends to work because moral goodness occasionally functions as a proxy for goodness in specific tasks.

Hence, the distinct societal role of trust is most obvious in situations characterized by an *epistemic asymmetry that obtains in virtue of task novelty*. In the above example, there is a crucial difference between what you ought to know (i.e., whether your partner is reliable at giving good haircuts) and what you actually know (i.e., they, a good person, have never cut her before), and this difference ought to serve as a counterweight when deciding whether to let her perform the action. By deciding to trust, one can attempt to compensate for this counterweight by appealing to the agent's moral trustworthiness in lieu of her task-relevant reliability. Continuing

[9] An astute observer will detect two kinds of reliability in play: one that fails to apply to the task at hand (here, cutting hair) and another which does apply to the agent's consistency in conforming to moral norms. Recall that trustworthiness is reliability plus an additional X- factor, so every trustworthy person is also reliable in the sense that she has reliability as an enduring character trait. However, this does not mean, of course, that every trustworthy person is reliable at all tasks.

[10] For readers familiar with Mayer et al.'s (1995) influential model of trust based on three dimensions (ability, integrity, and benevolence), this example is about the agent scoring near zero in ability but high in integrity and benevolence.

with the above example, you know your partner is a good person and so will do their best, and this, you believe, will be sufficient for them to produce an acceptable outcome. Intuitively, such a compensation strategy will reap societal benefits should it ultimately turn out to be justified (and which it will do, so long as, by and large, people are good people). Notice, however, that such cases of trust are not relevant for the machine learning applications in focus here, since they are always trained on a dataset about which there is a ground truth instrumental in establishing reliability.

One of the chief aims of the present discussion is to explicate how this epistemic asymmetry does not need to arise with certain machine learning applications, since these can be designed to be highly reliable at accomplishing tasks (and ought to be so designed). Thus, so long as certain features are in place (see below), there will be no gap between what you know (the AI is reliable at its tasks) and what you ought to know when interacting with the AI in these specific tasks. Also, it is our aim to make clear how the vulnerability associated with its deployment would not be alleviated if the AI were designed to be morally trustworthy. This is because the relevant vulnerability arises not from epistemic asymmetry, but from the possibility that the AI's prediction is wrong about the data it has been trained on, that is, from the possibility that the AI is unreliable.[11]

To sum up, our discussion unpacks what it means to say we should develop morally trustworthy AI. It means developing morally rational AI endowed with a responsiveness to moral norms. But as this discussion makes clear, such machines would not help mitigate the vulnerability that deploying such technology creates. To the extent that it does, it means developing the AI to offer various kinds of explanations for its decisions, capacities which do not entail moral rationality. Thus, it is wrongheaded to call for morally trustworthy AI. In the penultimate section below, we discuss how meeting the normative conditions of reliability can be done by developing quasi-metacognitive machines and how this strategy can be coupled with another one about designing them to be quasi-partners that offer explanations. These two features will mitigate vulnerability incurred by deploying them in medium to high-stakes hybrid decision-making environments.

4 Quasi-Metacognitive Machines and Alleviating the Problem of Vulnerability

Crucial for present purposes, calling for reliable AI, reliability understood as consistent competence, avoids the category mistake that skeptics of trustworthy AI are concerned about, since competence is completely orthogonal to morality (e.g., it is possible to be a perfectly competent liar). Thus, reliability, independent of trust, possesses unique normative conditions, conceptualized above as features of agents (see

[11] Considering the above interpretation of the significance of trustworthiness, one could interject that designing AI to be morally trustworthy would mean that we would be justified in deploying it in novel situations, which, in this context, would mean deploying it without the required training. To our knowledge, this would violate ethical codes as well as be simply a bad idea from an engineering standpoint, and so the point is somewhat moot, since its practical significance is unclear.

Section 2). Equipped with these notions of reliability and minimal agency, we can now ask about the capacities that AI needs to be consistently competent.

First, AI needs a capacity to process its decision as input; otherwise, it will have no way of monitoring its decision to ensure it exhibits consistency. Second, it needs to develop an internal model of its decisions, so that it can control its performance. Third, it requires a comparative function that contrasts performance to its internal model to determine whether decisions remain consistent. Fourth, this comparator function needs the capacity to produce an evaluation of performance that serves as the means by which the AI controls its behavior. Fifth, it needs a capacity to process this evaluation as input, so that the information carried by the evaluation can be used to update internal models and thereby control future performance. Finally, its decisions need to be a product of this complex process of monitoring, evaluating, and revising of internal models; otherwise, it will be unable to ensure competence in the event of error and in relevant counterfactual situations.

Interestingly, the architecture described above is analogous to, what has been called, "procedural metacognition", in which feedback signals inform the system implicitly about its own performance, signals which are used to update internal models and regulate the system's behavior (Proust, 2013). Procedural metacognition (also known as "system 1 metacognition", "evaluative metacognition", and "implicit metacognition") can be defined as, "the ability to evaluate the quality of one's own informational states, and the efficiency of one's own learning attempts, in order to regulate subsequent cognitive activities and behavior" (Goupil & Proust, 2023). Capacities for monitoring and evaluating one's own performance is a form of metacognitive sensitivity, and capacities for regulating behavior according to the evaluations of self-monitoring mechanisms is a form of metacognitive calibration and control. As such, one solution to the alignment problem is not developing machines to simulate moral rationality, but to simulate metacognition, and thus the goal of developing reliable AI is about developing quasi-metacognitive machines.[12]

As it turns out, a simulation of metacognitive architecture comes at little cost in many AI applications. In machine learning, what is known as the "F1-score" (also known as "model confidence") is a metric used to evaluate the performance of a classification model. It is the harmonic mean of precision and recall into a single score ranging from 0 to 1, which makes it useful for comparing the model's competence. Precision is the proportion of true positive results out of all predicted positive results (i.e., TP/(TP+FP)), while recall is the proportion of true positive results out of all actual positive results (i.e., TP/(TP+FN)). A higher F1-score indicates better performance, and it is routinely regarded as a measure of the confidence from the system in

[12] Why quasi? Bayne et al. (2019), a review article in Current Biology entitled "What is Cognition?", features answers to the titular question from 11 experts, ranging from philosophers of mind to cognitive neuroscientists and professors of artificial intelligence. The majority position is characterized by adaptive information processing, involving high degrees of flexibility, that enables stimulus-independence and causal reasoning. While the information processing of the AI systems for decision support admits of high degrees of flexibility, it is unclear how its processing is stimulus independent and produces genuine causal reasoning. Having said that, we deploy the term "*quasi*-metacognitive" to refer only to current AS/AI systems, which lacks stimulus independence and genuine causal reasoning. We do not wish to draw an a priori line on whether some future AI might manifest genuine cognition or even metacognition.

the generated classification or prediction. Moreover, several machine learning algorithms, such as logistic regression (which AFST employs) and Naïve Bayes, provide probability estimations for each class as part of their output. There are also Accuracy and Area Under the Curve scores. Thus, what ought to be called, "reliability scores" (rather than the anthropomorphic "confidence scores") inform about the reliability of the underlying process that determine the classification score.

In the context of experimental psychology, metacognitive competence refers to an individual's ability to accurately assess and evaluate their own cognitive processes, such as their memory, perception, or problem-solving abilities (Fleming & Lau, 2014). This is measured by contrasting metacognitive sensitivity to metacognitive bias. Metacognitive sensitivity refers to the ability to accurately detect and discriminate between one's own correct and incorrect judgments or performance, while metacognitive bias, on the other hand, refers to systematic deviations or distortions in metacognitive judgments, representing the tendency to make consistent errors in self-assessment, independent of one's actual performance or knowledge. While products of a fundamentally different mechanism, reliability scores can be the machine analogue to metacognitive sensitivity, with specific scores being analogues to calibrated confidence.

In other words, it would be advantageous for AFST to offer two scores. In addition to the already-present screening score about the likelihood that intervention is needed, there would also be a second score about the AI's reliability regarding the process that formed the screening score. At first glance, the reliability score might sound redundant, but comparing these two statements makes it clear that this is not the case: "I am 80% confident this case scores 18 out 20" versus "I am 20% confident this case scores 18 out of 20". Clearly, reliability scores will provide the screener with another useful piece of information to consider when deciding whether to rely on or deviate from the AI's recommendation.

This proposal is supported by recent empirical results. For example, Zhang et al. (2020) found that providing reliability scores not only led to the collective benefit of higher accuracy in discrimination and classification tasks, but it also facilitated human decision-makers in calibrating their reliance on the AI (i.e., mitigating the problem of vulnerability). This result is also supported by Rechkemmer and Yin (2022) that showed that providing reliability scores positively influences whether people believe an AI's predictions, so long as these scores are well-calibrated with performance (i.e., so long as the model exhibits consistent competence). Though the process by which reliability scores are calculated is vastly different from how confidence signals of procedural metacognition are produced, there are important functional similarities in terms of both. Both inform about the degree to which performance is sensitive to the relevant constraints, and both can be used to determine whether strategies are effective or whether ought to change.

As introduced above, many machine learning applications are already quasi-metacognitive in the manner described here, and so the demand of developing reliable AI is already met in many cases. What is left to understand is how to turn reliable AIs into reliable quasi-partners (only "quasi" because genuine partnership arguably requires joint responsibility; see Schmidt and Loidolt (2023) for a relevant taxonomy of partnerships), so that we might make sense of how it can be deployed to miti-

gate the problem of vulnerability for all relevant stakeholders. As a baseline enabling condition for partnering with AI (however simulated), it helps to recognize that AI advice is often constitutive of the final decision, such that the final decision ought to be regarded as a species of collective decision and the AI regarded as a contributor to this decision.

One plausible answer to the question of what enables a reliable agent to become a reliable partner is that the agent *communicates* her confidence to the other agents involved in the collective decision-making process. Not only is there a wealth of evidence that demonstrates that confidence communication incurs collective and collaborative benefits in collective decision-making (Bahrami et al., 2012), but there are conceptual reasons for linking the communication of confidence with the enabling conditions of a reliable partner. Recall from above that reliability is chiefly about consistent competence; indeed, having access to an agent's confidence clearly facilitates relying on them, since when confidence is well-calibrated, it serves as an approximation of the agent's competence. In other words, if I am confident that I can perform some action, then my confidence approximates the likelihood that I will be successful, and the inverse goes for when I am not confident.

As it regards providing explanations, our discussion above demonstrates that, rather than conceptual or emotional capacities (i.e., what it would *really take* to make AI morally trustworthy), it is the capacity to provide explanations that makes a meaningful contribution to mitigating the problem of vulnerability. In combination with the capacity to report reliability scores (i.e., the machine analogue to confidence), the AI should also be designed to provide explanations for the reliability scores. In the case of AFST, an explanation for a reliability score of 90 (out of 100) for a classification score of 18 (out of 20) might be, for instance, a parent's recent substance abuse charges. This would enable the AI to be a more reliable quasi-partner in the collective decision of whether to open an investigation. This is because, as our discussion has shown, confidence, when well-calibrated, is an indication of competence, and reliability, what we ought to be aiming toward, should be following competence.

Furthermore, explanations might be provided in counterfactual terms to facilitate the role of the AI as a quasi-partner. As the above normative analysis shows, reliability is also about competence in relevant counterfactual situations, and a responsiveness, when scoring, to counterfactual possibilities demonstrates this competence. For example, AFST might be designed to report that were the recent substance abuse charges absent in the case file, the reliability score would drop to 20% for a classification score of 18. Crucially, access to explanations of a counterfactual stripe is high on screeners' wishlist of AFST improvements (Kawakami et a., 2022), and counterfactual explanations have been shown to increase understanding of the reliability scores (Le et al., 2022). Moreover, the proposal is highly feasible since various machine learning methods exist for providing counterfactual descriptions (Erasmus et al., 2021).

5 Conclusion and Future Directions

The prevailing notion of developing morally trustworthy AI is fundamentally wrong-headed. What trustworthy AI ultimately amounts to is morally rational AI, which, feasibility aside, is unnecessary for promoting optimal calibration of the human decision-maker on AI assistance. Instead, reliability, exclusive of trust, ought to be the appropriate goal of ethical AI, as it entails the normative constraints that, when met, are effective in mitigating the vulnerability created by having AI make meaningful contributions in high-stakes decision-making environments. These normative demands can be met by developing AI to exhibit quasi-metacognitive competence as quasi-partners, which is largely already in place in machine learning systems. Crucially, the AI ought to be designed to report reliability scores to the human decision-makers due to how such scores enable the human to calibrate her reliance on the AI's advice. Moreover, these scores could be more effective if accompanied by reasons and counterfactual explanations. By highlighting the importance of such explanations in promoting optimal calibration, without any need for moral trust, we have pointed the way to developing more ethical applications by way of developing quasi-metacognitive machines.

Acknowledgements We would like to thank the anonymous reviewer for their helpful and valuable feedback that greatly improved the quality of this manuscript.

Author Contributions The first draft and peer review revisions of the manuscript were written by John Dorsch, and Ophelia Deroy commented on drafts of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by Bayerisches Forschungsinstitut für Digitale Transformation (bidt) Co-Learn Award Number KON-21-0000048 and HORIZON EUROPE European Innovation Council (EIC) EMERGE Grant agreement ID: 101070918.
Open Access funding enabled and organized by Projekt DEAL.

Data Availability Not Applicable.

Declarations

Ethics Approval and Consent to Participate Not Applicable.

Consent for Publication We hereby give our consent for the submitted manuscript to be published in the journal Philosophy & Technology and bestow onto the publisher all rights that come with this.

Competing Interests The authors declare that they have no competing interests.

References

Alvarado, R. (2023). What kind of trust does AI deserve, if any? *AI and Ethics*, *3*(4), 1169–1183. https://doi.org/10.1007/s43681-022-00224-x.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 3–8. https://doi.org/10.1037/a0025708.

Baier, A. (1986). Trust and Antitrust. *Ethics*, *96*(2), 231–260. https://doi.org/10.1086/292745.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2021). Is the most accurate AI the best teammate? Optimizing AI for Teamwork. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(13), 11405–11414. https://doi.org/10.1609/aaai.v35i13.17359.

Bayne, T., Brainard, D., Byrne, R. W., Chittka, L., Clayton, N., Heyes, C., Mather, J., Ölveczky, B., Shadlen, M., Suddendorf, T., & Webb, B. (2019). What is cognition? *Current Biology*, *29*(13), R608–r615. https://doi.org/10.1016/j.cub.2019.05.044.

Bryson, J. (2018). AI & global governance: no one should trust AI. *United Nations Centre for Policy Research*. Retrieved April, 27, 2023: https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html#:~:text=We%20should%20focus%20on%20AI,of%20our%20institutions%20and%20ourselves.

Buechner, J., & Tavani, H. T. (2011). Trust and multi-agent systems: Applying the diffuse, default model of trust to experiments involving artificial agents. *Ethics and Information Technology*, *13*(1), 39–51. https://doi.org/10.1007/s10676-010-9249-z.

Cherkowski, S., Walker, K. D., & Kutsyuruba, B. (2015). Principals' Moral Agency and ethical Decision-Making: Toward a transformational Ethics. *International Journal of Education Policy and Leadership*, *10*(5), n5. https://eric.ed.gov/?id=EJ1138586.

Christian, B. (2020). *The Alignment Problem: Machine learning and human values*. WW Norton & Company.

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology*, *14*(1), 53–60. https://doi.org/10.1007/s10676-011-9279-1.

Davidson, D. (1963). Actions, reasons and causes. *Journal of Philosophy*, *60*, 685–670.

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376638

Deroy, O. (2023). The Ethics of Terminology: Can we use human terms to describe AI? *Topoi*, *42*(3), 881–889. https://doi.org/10.1007/s11245-023-09934-1.

Deutsch, M. (1977). *The resolution of conflict: Constructive and destructive processes*. Yale University Press. https://doi.org/10.12987/9780300159356.

Dretske, F. (1988). *Explaining Behavior*. MIT Press.

Erasmus, A., Brunet, T. D. P., & Fisher, E. (2021). What is Interpretability? *Philosophy & Technology*, *34*(4), 833–862. https://doi.org/10.1007/s13347-020-00435-2.

European Commission (2019). *Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee and the Committee of the Regions*, Building trust in human-centric artificial intelligence. COM(2019) 168 final (8 April 2019).

Evans, G. (1982). The Varieties of Reference. Oxford: Oxford University Press.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2014.00443. 8.

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.

Fodor, J. (2008). *LOT2: The language of thought revisited*. Oxford University Press.

Gibbard, A. (1990). *Wise choices, apt feelings: A theory of normative Judgment*. Harvard University Press.

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057.

Goldman, A. I. (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy*, *73*(20), 771–791. https://doi.org/10.2307/2025679.

Goupil, L., & Proust, J. (2023). Curiosity as a metacognitive feeling. *Cognition*, *231*, 105325. https://doi.org/10.1016/j.cognition.2022.105325.

Green, B., & Chen, Y. (2019). The principles and limits of Algorithm-in-the-Loop decision making. *Proceedings of ACM Human Computer Interactions*, *3*(CSCW), Article50. https://doi.org/10.1145/3359152.

Hardin, R. (2002). *Trust and Trustworthiness*. Russell Sage Foundation.

Haugeland, J. (1990). The Intentionality All-Stars. *Philosophical Perspectives*, *4*, 383–427. https://doi.org/10.2307/2214199.

Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing ourselves together: The cultural origins of metacognition. *Trends in Cognitive Sciences*, *24*, 349–362. https://doi.org/10.1016/j.tics.2020.02.007.

Hieronymi, P. (2004). The Force and Fairness of blame. *Philosophical Perspectives*, *18*(1), 115–148. https://doi.org/10.1111/j.1520-8583.2004.00023.x.

Johnson, D. (2006). Computer systems: Moral entities but not moral agents. *Ethics of Information Technology*, *8*, 195–204. https://doi.org/10.1007/s10676-006-9111-5.

Kamar, E. (2016). Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, New York, 9–15 July 2016, 4070–4073.

Kawakami, A., Sivaraman, V., Cheng, H. F., Stapleton, L., Cheng, Y., Qing, D., Perer, A., Wu, Z. S., Zhu, H., & Holstein, K. (2022). Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, New Orleans, LA, USA. https://doi.org/10.1145/3491102.3517439.

Lahno, B. (2001). On the emotional character of Trust. *Ethical Theory and Moral Practice*, *4*(2), 171–189. https://doi.org/10.1023/A:1011425102875.

Le, T., Miller, T., Singh, R., & Sonenberg, L. (2022). Improving model understanding and trust with counterfactual explanations of Model confidence. *arXiv Preprint arXiv*:220602790.

Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLOS Computational Biology*, *15*(4), e1006973. https://doi.org/10.1371/journal.pcbi.1006973.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

Longin, L., Bahrami, B., & Deroy, O. (2023). Intelligence brings responsibility - even smart AI-assistants are held responsible. *iScience*, 107494. https://doi.org/10.1016/j.isci.2023.107494.

Luhmann, N. (1979). *Trust and Power*. Wiley.

Mainz, J. T. (2023). Medical AI: Is trust really the issue? *Journal of Medical Ethics*. https://doi.org/10.1136/jme-2023-109414.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792.

McDowell, J. (1994). *Mind and world*. Harvard University Press.

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, *21*(4), 18–21. https://doi.org/10.1109/MIS.2006.80.

National Academies of Sciences. (2017). Communicating Science effectively: A Research Agenda. *National Academies Press*. https://doi.org/10.17226/23674.

O'Connor, C., & Weatherall, J. O. (2019). *The misinformation age: How false beliefs spread*. Yale University Press.

Peacocke, C. (1983). Sense and Content: Experience, Thought and Their Relations. Oxford: Oxford University Press.

Peacocke, C. (1992). *A study of concepts*. The MIT.

Porsdam Mann, S., Earp, B. D., Nyholm, S., Danaher, J., Møller, N., Bowman-Smart, H., Hatherley, J., Koplin, J., Plozza, M., Rodger, D., Treit, P. V., Renard, G., McMillan, J., & Savulescu, J. (2023). Generative AI entails a credit–blame asymmetry. *Nature Machine Intelligence*, *5*(5), 472–475. https://doi.org/10.1038/s42256-023-00653-1.

Proust, J. (2013). *The philosophy of Metacognition: Mental Agency and Self-Awareness*. Oxford University Press.

Rechkemmer, A., & Yin, M. (2022). When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, New Orleans, LA, USA. https://doi.org/10.1145/3491102.3501967.

Ryan, M. (2020). In AI we trust: Ethics, Artificial Intelligence, and reliability. *Science and Engineering Ethics*, *26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y.

Samant, A., Horowitz, A., Xu, K., & Beiers, S. (2021). Family surveillance by algorithm. *American Civil Liberties Union*. https://www.aclu.org/fact-sheet/family-surveillance-algorithm

Scarantino, A. (2018). & de Sousa. R. Emotion. *The Stanford Encyclopedia of Philosophy (Summer 2021 Edition)*. Edward N. Zalta (Ed.). https://plato.stanford.edu/archives/sum2021/entries/emotion.

Schmidt, P., & Loidolt, S. (2023). Interacting with machines: Can an Artificially Intelligent Agent be a Partner? *Philosophy & Technology*, *36*(3), 55. https://doi.org/10.1007/s13347-023-00656-1.

Sharkey, A. (2020). Can we Program or Train Robots to be good? *Ethics and Information Technology*, *22*(4), 283–295. https://doi.org/10.1007/s10676-017-9425-5.

Sheridan, T. B., & Hennessy, R. T. (1984). *Research and modeling of supervisory control behavior. Report of a workshop*. National Research Council Washington DC Committee on Human Factors.

Simon, J. (Ed.). (2020). *The Routledge Handbook of Trust and Philosophy* (1st ed.). Routledge. https://doi.org/10.4324/9781315542294.

Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy*, *48*, 187–211.

Tsakiris, M., & De Preester, H. (Eds.). (2018). *The interoceptive mind: From homeostasis to awareness*. Oxford University Press.

Vaithianathan, R., Putnam-Hornstein, E., Jiang, N., Nand, P., & Maloney, T. (2017). *Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation*. Center for Social data Analytics.

Wu, W. (2013). Mental Action and the threat of Automaticity. In A. Clark, J. Kiverstein, & T. Vierkant (Eds.), *Decomposing the Will* (pp. 244–261). Oxford University Press.

Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain. https://doi.org/10.1145/3351095.3372852.