Topoi (2023) 42:881–889
https://doi.org/10.1007/s11245-023-09934-1

The Ethics of Terminology: Can We Use Human Terms to Describe AI?

Ophelia Deroy[1,2,3]

Accepted: 19 May 2023 / Published online: 8 June 2023

Abstract
Despite facing significant criticism for assigning human-like characteristics to artificial intelligence, phrases like "trustworthy AI" are still commonly used in official documents and ethical guidelines. It is essential to consider why institutions continue to use these phrases, even though they are controversial. This article critically evaluates various reasons for using these terms, including ontological, legal, communicative, and psychological arguments. All these justifications share the common feature of trying to justify the official use of terms like "trustworthy AI" by appealing to the need to reflect pre-existing facts, be it the ontological status, ways of representing AI or legal categories. The article challenges the justifications for these linguistic practices observed in the field of AI ethics and AI science communication. In particular, it takes aim at two main arguments. The first is the notion that ethical discourse can move forward without the need for philosophical clarification, bypassing existing debates. The second justification argues that it's acceptable to use anthropomorphic terms because they are consistent with the common concepts of AI held by non-experts—exaggerating this time the existing evidence and ignoring the possibility that folk beliefs about AI are not consistent and come closer to semi-propositional beliefs. The article sounds a strong warning against the use of human-centric language when discussing AI, both in terms of principle and the potential consequences. It argues that the use of such terminology risks shaping public opinion in ways that could have negative outcomes.

Keywords Trustworthy AI · Anthropormorphism · Animism · Legal Person · Naive concepts · AI

1 Introduction

Trustworthy Artificial Intelligence (AI) has become the flagship for ethicists and governmental agencies. When the High-Level Expert Group on Artificial Intelligence of the European Union released its first ethics guidelines in 2018, it did so intending to "put forward a set of 7 key requirements that AI systems should meet in order to be deemed trustworthy".[1] Professional ethicists approved the guidelines and the label (e.g. Floridi 2019). The Organisation for Economic Cooperation and Development (OECD) followed suit, using the same term to issue its recommendations about using AI for education.[2] When Microsoft decided to fund its ethics project, it did so under the same "Trustworthy AI" tagline.[3] Searching scholarly articles using the expression already returns more than 93,000 publications.

The repetition of the term raises concern: if the word is echoed across governments, industries, and universities, will citizens increasingly believe that AI can be trusted as an effect of exposure only? Repetition is an effective rhetorical device: the same sentence is processed more fluently as it gets seen multiple times, and as a result, gets more likely to look true, independently of its initial plausibility (see Dechêne et al. 2010, for a review). Still, the main concern with the expression "trustworthy AI" is not this possible illusory truth effect. Applying the concepts of trust and trustworthiness to artificial intelligence encourages us treat it not as a mere tool but as a system capable of human-like

✉ Ophelia Deroy
ophelia.deroy@lmu.de

[1] Faculty of Philosophy, Munich Center for Neuroscience, Ludwig Maximilian University, Geschwister Scholl Platz 1, 80539 Munich, Germany

[2] Munich Center for Neuroscience, Ludwig Maximilian University, Munich, Germany

[3] School of Advanced Study, Institute of Philosophy, University of London, London, UK

[1] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (2019).

[2] https://www.oecd.org/education/trustworthy-artificial-intelligence-in-education.pdf (2020).

[3] https://www.microsoft.com/en-us/research/project/trustworthy-ai/.

characteristics. It is this encouragement that this articles intends to examine.

2 Preliminary Considerations

We sometimes say things like "I trust my car" or "I trust this brand"—so what's wrong with saying "I trust my voice assistant" or "I sometimes trust what chatGPT says"? The problem here is that using trust is a loose use of the term. For those who do not share the intuition that using the word "trust" for a car or inanimate object is odd, a stronger intuition may be prompted in the case of the claim "the car is trustworthy".

In a stricter reading, trust normatively implies more than an expectation of reliability: it also captures other components, like honesty, benevolence, or compliance with social and moral norms which only humans can possess. Our concept of trust is canonically shaped around trust between humans which requires more than simply relying on someone to perform what we trust them to do correctly. Reliance, in other words, is part of trust but not enough (Golberg 2020).

From a philosophical standpoint, trust needs to be based on the minimal assumption that the other will consent to do what we trust them to do: Trust does not act by coercing but by a willingness to comply. Trust is not needed if compliance is mechanical and simply given. As Annette Baier writes, "trusting can be betrayed, or at least let down, and not just disappointed" (1986, p. 235). Going one step further, an action dictated by trust seems to engage a reciprocal commitment to do the right thing for the other party, which can be described as moral or altruistic.

The existence of a human concept of trust, with stricter demands, does not mean that the term trust is not sometimes used to mean "mere reliance", but that it at least carries with it an ambiguity when used about AI. Law and Scheutz propose to analyse this ambiguity as separating "performance-based trust" versus "relation-based trust". "Performance-based trust", the authors note "centers around the robot being trusted to be reliable, capable, and competent at its task or tasks, without needing to be monitored by a human supervisor. A performance-based trust may also depend on the robot's transparency, responsiveness, and predictability. Relation-based trust, on the other hand, implies that a robot is trusted as a social agent. A person with whom it interacts can be vulnerable emotionally and may trust that the robot will be sincere and ethical. Relation-based trust means that a person trusts the robot to be part of society in some way, not just off in a factory doing a job without any expectation of knowledge of social norms". (Law and Scheutz 2021, p. 28). When I say that I trust science, I may be ambiguous between expecting that science is reliable, and expecting that scientists and scientific institutions are responsible, moral beings. When I say that I trust my car, I probably simply mean that I rely on it, or I anthropomorphise the car. But what about committees of experts using the word "trust" to speak about AI?

The ambiguity of the words "trust" and "trustworthy" should be identified and eliminated in the process of a critical expert discussion. But can this mean that experts have decided that the right category to communicate about AI in legal or ethical frameworks is a category where AI is sufficiently similar to humans to be capable of "reliation-based trust"?

Several commentators and researchers see the use of anthropomorphic terms to communicate about AI as strategic misinformation or marketing rhetoric by industries.[4] "When news articles uncritically repeat PR statements, overuse images of robots, attribute agency to AI tools, or downplay their limitations, they mislead and misinform readers about the potential and limitations of AI", Sayash Kapoor and Arvind Narayanan wrote in a checklist of AI reporting pitfalls posted online. "When we talk about AI", Kapoor says, "we tend to say things like 'AI is doing X—artificial intelligence is grading your homework', for instance. We do not talk about any other technology this way—we do not say, 'The truck is driving on the road', or 'a telescope is looking at a star'. It is illuminating to think about why we consider AI different from other tools. In reality, it is just another tool for doing a task".[5]

Ethics committees and policy experts should also be different from marketing people: While experts are held to exacting standards of clarity and should avoid loose or figurative language, it is unlikely that they would resort to strategic misinformation akin to typical PR efforts for artificial intelligence. So could there be another justification for using the word "trustworthiness" to communicate about AI? How good is this justification, and, as a test, how strongly does it recommend using "trustworthiness" instead of "reliability"?

Before delving into these inquiries, it is necessary to outline three initial observations. Firstly, our analysis centers on characterizing AI as "trustworthy", premised on the idea that this label is typically associated with human agents and implies certain moral or social sensitivities and intentions. Although this assumption is widely recognized, it remains subject to debate. In particular, we need to recognise that many experimental studies use the

[4] https://www.theguardian.com/commentisfree/2019/jan/13/dont-believe-the-hype-media-are-selling-us-an-ai-fantasy. See also Crawford (2021). The ultimate motives for using the rhetoric of trust rather than reliability to deceive the public are unclear.

[5] https://www.latimes.com/business/story/2022-10-07/artificial-intelligence-ai-hype.

idea that human users "trust" AI or robots as an accepted frame to measure the attitudes of human users (though the measures themselves are highly heterogeneous and not all validated, see Law and Scheutz 2021; Perrig et al. 2023). Although the term "trustworthy AI" is employed as an example of human characteristics that can be ascribed to AI, it is not the only one. Therefore, the arguments presented in this paper can be adjusted to apply to other attributes, particularly for those who are skeptical about whether "trustworthiness" entails human or intentional prerequisites.

Second, focusing on human-specific terms like "trustworthiness" suggests that the problem is upgrading AI to a human or human-like status. But what about other problematic terms not reserved for humans and still imply that AI is more than a machine? Some agentive terms apply to humans and non-human animals and are also problematic. For instance, we say or hear that "The AI system chose the right solution", "the autonomous car swerved", or "The robot decided to lift the cup". Following existing accounts such as Tomasello (2022), we agree that agency is a plural concept that can extend to entities showing minimal forms of goal-directed behaviour and more complex ones showing more complex forms of intentional, rational or even moral control. While attributions of moral agency to AI fall within attributions of human characteristics to AI and therefore fall within our current concerns, it could be that other forms of agency attributions, particularly minimal attributions of goal-directed behaviour, are more grey areas and will not be primarily addressed here.

Finally, the primary target here is expert committees and public institutions, not companies or journalists. Companies may have commercial or persuasive reasons to use specific terms. Governmental agencies and national or international public ethics committees should be held accountable to higher normative standards regarding communication, which include nonpartisan norms such as fairness, impartiality, neutrality, and measured choice. The proper list of such standards is not fixed in stone, and—as we will see—it involves trade-offs between norms of accuracy and norms serving other social goods. What is clear is that public ethics committees and governmental agencies should not use communicative means that amount to malevolent deception or serve the private interests of a few over the common good. Journalism is a grey area, which, although bound by norms of communication, has its own right to personal interpretations and rhetorical license. Because of this complexity, it will not be part of the current argument.

3 The Importance of an Ethics of AI Science Communication

The central goal here is in the nascent field of the ethics and politics of science communication (e.g. Medvecky and Leach 2019) which looks at how science and technology should be presented and represented to citizens by public institutions and administrations, especially those composed of appointed experts. The reports of expert panels and ethics committees, in particular, when dealing with the use of technology, need not exclusively aim to perform science communication for them to have still to rely on elements of science communication in characterising the technology and science it is based on.

As an example, neuroscientists have objected that the use of the concept of 'fetal pain' in the new US regulations against abortion rights is factually incorrect: as the authors note, "Abortion policy has profound moral and ethical consequences and therefore needs to be grounded in the most accurate scientific arguments, as well as a clear understanding of what we mean when we use the term pain" (Solomons and Ianetti 2022). At the same time, the authors note that the use of the term is all the more important than it is used to communicate facts to the public. Similarly, ethical and institutional texts should be responsible for upholding good practice in communicating scientifically relevant evidence or when communicating about scientific and technological issues. For our present purposes, "communication" should be understood as referring only to such official communication. With these cautionary remarks and clarification in mind, we can return to the core question.

4 The Argument of Ontological Match

Could something make the official terminology of trustworthiness appropriate in the case of AI? An obvious justification would be that AI is, metaphysically, a human or a human-like agentive entity or is likely to be in the foreseeable future. Call these justifications that of "actual ontological match" or "forecasted ontological match". They can be formulated as follows:

> Actual ontological match: using the word W, referring to the property P, is appropriate to communicate about AI if and only if AI systems really possess the property P or are the type of entity which possesses P.
>
> Forecasted ontological match: the use of the word W, referring to the property P, is appropriate to com-

municate about AI if and only if AI systems can be reasonably predicted to possess the property P or to be the type of entity which possesses P in the foreseeable future,

For instance, in the case of trustworthiness, supposing that it only applies to entities which can have benevolent intentions or dispositions to understand social or moral norms, the argument would mean that AI systems either can have such intentions or dispositions or can be reasonably predicted to possess them in the foreseeable future.

It is fair to say that the justification by actual ontological match stands on fragile ground (e.g. see the discussions in Dennett 2019; Nyholms 2023; Popa 2021; Veliz 2021; as well as the older discussions stemming from Searle 1980). While some philosophers like Daniel Dennett firmly consider that AI qualifies as inanimate tools, there is little philosophical and scientific agreement that AI can have the intentional, social or moral capacities that make them comparable or analogue to humans and properly "trustworthy". A statement claiming or implying they do is at least unjustified and likely false. As far as the norm of accuracy is concerned, an institution cannot justify communicating about AI in these terms by saying that the expression is accurate.

If the justification from an actual ontological match is not good, does the forecasted match fare better? The question hinges on what one reads in the expressions "reasonable forecast" and "foreseeable" future. Despite their indeterminacy, the terms should exclude highly implausible scenarios based on mere imaginative projections and disconnected from actual scientific evidence. They should also exclude scenarios that could hold in 50 years or more—considered the upper range for predictions and taps into mere speculation. In other words, the justification by "forecasted ontological match" can not collapse into long-termist arguments considering speculative future scenarios that go well beyond the 50-year range, as those cannot be assigned a reasonable probability.

Forecasts represent specific probabilistic judgements about the future, which allow governments and public institutions to form more to less reliable predictions about future events (e,g. Tetlock and Gardner 2015). The empirical literature on political and economic forecasting shows that specific steps, such as aggregation, selection, training and discussion, enable us to reach reliable forecasts and minimise what is known as prediction error (e.g. Dhami and Mandel 2021, Dezecache et al. 2022; Ferreiro et al. 2023).

Returning to the subject of AI, there currently exists no rigorously conducted and published forecasting exercise regarding the properties that AI is likely to acquire in the near future. There is no consensus regarding any such forecast for the next twenty years, and the cautionary prediction is that AI will continue to lack human-like characteristics, as previously discussed. Furthermore, as the timeframe of a forecast extends further into the future, its reliability decreases. Thus, even if a forecast for 20 years or more assigned a significant probability to AI possessing human-like properties, such a forecast would remain unreliable. In the absence of reliable forecasts that AI will possess human-like properties, the argument based on "forecasted match" provides a weak justification for utilizing anthropomorphic attributes such as "trustworthiness" when communicating about AI.

In conclusion, the notion of an actual or projected ontological match presents a tenuous foundation for justifying the application of terms such as "trustworthiness" to AI. This does not imply a dismissal of the current philosophical perspectives that support the attribution of certain, typically derivative or rudimentary, forms of intentionality, agency, or other human-specific qualities to AI. Rather, it highlights that such perspectives remain part of an ongoing academic discourse and have yet to gain sufficient acceptance as the correct ontology to warrant a shift in public discourse. The focus here is not on debating the merits of various metaphysical viewpoints on AI but rather on determining the appropriate ontological categories to utilize in public discourse.

5 The Argument of Communicative Efficacy and fit with folk Conceptions

5.1 Social Perception Matters

Another defense for utilizing expressions such as "trustworthy AI" is that they align not with the established scientific ontology, but with the "naive" categories commonly used by citizens and users. In other words, the rationale behind legislating for trustworthy AI would be that individuals perceive AI as human-like entities capable of being trusted. From a normative standpoint, policies and ethical recommendations must account for how users relate to these technologies and regulate the interactions between AI producers and users based on culturally or socially accepted terms. Cockelberg (2011), for instance, argues that societal norms and perceptions take precedence over ontology when it comes to the ethical and policy aspects surrounding AI. An example of this would be that if users are inclined to trust a chatbot or caregiver robot, then laws and ethical policies should utilize this category and assess how the AI system fulfills this expectation.

In various contexts, the argument of "cultural match" is employed without explicit acknowledgment. For example, although wine and alcohol are scientifically no different from other addictive substances, they are not commonly regarded as drugs in many parts of the world. The popular concept of "drugs" encompasses cocaine, heroin, opioids,

and cannabis, but excludes beer, wine, or whisky. Consequently, official discourse often adopts this cultural understanding and employs different language when referring to alcohol compared to heroin or even cannabis. However, in both the cases of AI/human and alcohol/drugs, the point is not to assert that categorical distinctions are grounded in metaphysical differences (which is highly disputable). Another example is to use, for instance, the word mushroom to communicate about mycelia if the category is familiar to people and if people are not familiar with the difference or with technical names. The point is a pragmatic argument for "cultural match", whereby using such conceptual boundaries renders official public discourse comprehensible or acceptable to citizens. (See Hoffman 2014 for a similar argument that legal texts are more effective when they align with citizens' intuitions.)

In conclusion, the argument of "cultural match" has some initial support in recommended practices of science and communication. It may be justifiable to use familiar categories, even if imprecise, to communicate about complex technical concepts if it makes the communication more efficient and benefits the recipients. (see John 2018, for a related discussion of science communication ethics). Of course, some degree of precision is lost in doing so. Yet, the switch can still be justified if the communication is made more efficient and the end result of the communication benefits the recipients.

However, from a normative perspective, there is still the question of how to assess if these criteria are met when official communication adopts a culturally accepted category that is not scientifically sound. In the case of AI, this argument also needs further justification from a descriptive perspective, as it only holds if lay users and citizens treat AI as humans or sufficiently human-like and trust them in the same way they would trust another human. Therefore, the question of whether the use of terms like "trustworthy AI" corresponds to the way people view and interact with AI needs to be examined.

5.2 Descriptive fit: Do People Anthropomorphise AI?

How do people commonly think about AI? Do they spontaneously use the same concepts or categories for artificially intelligent systems as the ones they use for humans? The question is empirically complex, not just because of the variety of AI systems that exist but also because the way we commonly think about other human minds is a vast area.

When it comes to human-looking robots, behavioural studies suggest that people interact with such AI-driven entities as if they were human agents rather than mere machines—at least if they look sufficiently human (for an overview, see Broadbent 2017). For instance, people apply stereotypical social categories such as gender (e.g., Eyssel and Hegel 2012) or (out)group memberships (Eyssel and Kuchenbrandt 2012; Kuchenbrandt et al. 2013) to robots. They also display typical social behaviour toward them: they punish a robot that admits wrongdoing (Lee et al. 2021) and accept its apologies (Lee et al. 2010). Further on the moral side, evidence shows we tend to recognise AI as partly accountable for their mistakes (Kahn et al. 2012b). Such behaviour suggests that humans attribute human traits to robots.

Various studies suggest that people attribute intentionality to AI agents (for an overview, see Perez-Osorio and Wykowska 2020). Thellman et al. (2017) found that people even ascribe the same level of intentionality to humanoid robots and human agents when asked to rate it from different images and verbal descriptions of their actions and decisions. Graaf and Malle (2019) also gave people verbal descriptions of robot and human behaviours across different contexts and asked them to explain why the agent had performed them. People used the same conceptual toolbox of behavioural explanations for human and robot agents. However, an equally important number of studies using interactive set-ups suggest that naive users draw a difference between AI and humans: they notably trust AI less in some conditions (Burton et al. 2020), will reciprocate less towards an AI than a human stranger (Karpus et al. 2021), and stop being cooperative when they know that another agent is a bot (Ishowo-Oloko et al. 2019).

Following Gray et al. (2007), Geiselmann et al. (2023) suggest that the attributions of human characteristics to AI should not confuse the attribution of a capacity for agency and the attribution of a capacity for experience (such as hunger, fear, pain, etc.). For instance, lay people attribute a high degree of experience to a baby but no agency; they also attribute high agency but no experience to a deity. To determine if robots are considered human agents, it is essential to assess the degree to which both agency and experience are ascribed to them. While evidence that AI is granted the explicit ability to plan and act and elicit the same action representation mechanisms as human interaction partners is strong (e.g. Chaminade et al. 2012; Bisio et al. 2014), the same is not true for the ability to experience or have other mental states. Research has found that people are quicker to interpret a human's gaze or predict their actions compared to humanoid robots. This advantage in processing human gaze was observed in tasks that require representing others' minds, implying that people are less able to infer the future actions of robots than humans (Tidoni et al. 2022). However, people can still extract non-mentalistic information from a robot's gaze.

Neuro-imagery studies confirm that mechanisms linked to the attribution of sentience or mental states are not fully activated or are less activated when interacting with humanoid

robots. For example, increased neural activity was observed in mentalising areas (such as the temporoparietal junction and dorsal prefrontal cortex) during human-human interactions but not human-robot interactions. This was observed during eye contact (Kelley et al. 2021) and in conversation (Rauchbauer et al. 2019). Even when AI is provided "a human face", viewing robotic facial expressions evokes less activity in mentalising areas than viewing human facial expressions (Hmamouche et al. 2020).

5.3 Three Interpretations

The evidence so far opens more than solves questions: If trust requires features like benevolence or vulnerability, then folk concepts of AI do not attribute such features to AI. The debate, however, can be transferred to a more general level: Does evidence show that people commonly perceive AI as different from humans in degrees but not in kind? Do they, in other words, accept to extend the category of "human" to them as if they are extreme or non-prototypical cases, but cases nonetheless? Or are the differences they make between AI and humans sufficient to suggest that they have started to create a new ontological category for AI, as suggested by Kahn et al. (2017)?

Here, Deroy (2021) has recommended a third interpretation, distinct from the "new ontological category" and the "extended human-category" interpretations: The idea that people's category for.

AI is inherently part of semi-propositional beliefs, which do not fully integrate with other beliefs we may have, either about humans or machines. The naive concept, in this sense, is closer to positing a "ghost in the machine", which brings about an interesting similarity with cognitive and philosophical accounts of religious and spiritual beliefs. What Sperber means by semi-propositional or half-understood beliefs (Sperber 1982, 1997) is beliefs where the "content is not just vague; it is mysterious to the believers themselves and open to an endless variety of exegeses" (Sperber 2009, p. 534).

Appealing to semi-propositional beliefs helps reconcile otherwise problematic sets of attitudes: Take animists, for instance, who may accept that "dead spirits are watching the living" and that "eyes are necessary to watch" and still agree that "dead spirits have no eyes". The three beliefs are logically inconsistent, as they cannot all be true. However, instead of considering that the person holding these beliefs is irrational and self-contradictory, it is possible to consider that one of the beliefs ("dead spirits are watching the living") is a semi-propositional belief: it reflects the belief that this is the right belief to have, or the right way of speaking about dead-spirits but does not mean that the believer takes it as having to integrate with other beliefs.[6] Humans thinking about AI may equally accept that "the driverless car made a mistake", "intentions are necessary to count a failure as a mistake", and "driverless cars do not really have intentions" because they hold semi-propositional beliefs about AI. The account fits well with the fact that most of what lay users know about AI is, de facto, said to us by others: industries, media, and artists provide them with various, not all consistent, pieces of information about AI, which they then believe to be the right way to think and speak about AI—but do not necessarily integrate with other beliefs.

5.4 Normative Question: Is it Right to Respect Naive's Views?

There are multiple ways to interpret the mixed evidence regarding people's naive categories of AI. Three views stand out: the extension view, which posits that naive users extend their category of humans to include AI, even though they do not view AI as central or prototypical in that category; the novelty view, which suggests that naive users have a new category for AI that shares some features with the category of humans but is distinct; and the semi-propositional view, which proposes that naive users hold a non-fully propositional set of beliefs about AI that do not form a consistent concept, reminiscent of elements of religious beliefs, such as ghosts or spirits.

Each view recommends something different regarding using human terms to communicate about AI. The extension view is the one that makes this use the most legitimate, along the following lines: because people tend to treat AI as humans, communicating about AI using human terms (like trustworthiness) is most likely to be relevant and understandable to people's decisions about AI. On the other hand, the novelty and semi-propositional views do not give the green light to use human terms in such an easy way, but for different reasons. According to the novelty view, using human terms to communicate about AI will likely result in a category mistake—even by taking naive categories as the standard. For example, suppose the new category formed for AI is as distinct from the category reserved for humans as it is from the category used for inanimate machines. In that case, there is also no good reason to think it is more appropriate to stretch the language and borrow terms from the neighbour category of humans rather than machines. In other words, the novelty view leaves the choice of talks of "trustworthy AI" looking at best arbitrary—and at worst, motivated by the preferences or values of the communicator.

The semi-propositional view issues a different verdict. Using human terms to communicate about AI contributes to feeding a fundamentally non-logical set of beliefs, which will not integrate with other beliefs about the world. Similarly, talking about ghosts or supernatural entities and

[6] In other words, it counts as a meta-belief.

attributing them properties or powers that pertain to humans (e.g. "ghosts can see us") instils the idea that this is the right way to speak and think about ghosts, though accepting such propositions does not mean that one accepts their logical consequences ("eyes are necessary to see; therefore ghosts have eyes"). As a consequence, the semi-propositional view issues a somewhat different verdict regarding the use of human terms to speak about AI: on the one hand, it converges with the extended view in seeing the use of human terms as compatible with the current psychological treatment of AI, making it a possibly efficient way to speak about AI, but on the other hand, it also sends a strong warning against reinforcing—or even creating—an isolated set of semi-propositional beliefs about AI, that will not cohere with other beliefs. To put it more crudely, it would be like creating a form of religious, non-rational way of talking about AI. This way of talking may be successful, as the human mind is at ease with such semi-propositional beliefs but does not favour the right type of rational attitude towards AI.

For now, it suffices to stress that using human terms to describe AI would only be legitimate, according to a communicative argument, if we had good evidence that naive users categorise AI as humans. So far, this evidence is missing and at least insufficient.

6 The Legal Argument: Extending Legal Personhood to Inanimate Entities

So if ontological adequacy, actual or predicted, cannot justify that ethics committees use concepts used for human agents to refer to AI, what other justifications are there? One interesting line of argument arises from the legal side. Existing legislative and ethical systems usually use legal persons to assign rights, duties and sanctions. Moreover, the category of a legal person is already extended to non-human entities like corporations: In other words, the law offers a precedent where entities which are not literally human, or even sentient, are granted human-like characteristics, such as being a bearer of responsibilities, duties, or capable of commitments. In the same way a corporate could find itself in a breach of trust, could an AI also be trusted and breach trust?

Legal personhood is recognised for various entities and is a debated issue, especially for corporations. Several people have discussed its promises and limits for AI, at least since Solum (1992), and their arguments come down to three main families.

The main argument for treating AI as a legal person revolves around comparing corporations and AI entities. Dan-Cohen (2016) argues, for instance, that corporations and AI entities share the characteristics of organisations, such as temporal independence and the ability to exist beyond their creators, as well as having complex and formal structures with organisational intelligence. However, Solaiman (2017) disagrees, stating that machines do not possess the necessary legal recognition and that the comparison between corporate personhood and AI personhood ignores that corporations are symbolic representations of people.

However, two other families of arguments are provided, one on pragmatic grounds and the other on epistemic ones. Some, like Novelli (2022), argue, for instance, that the category of legal person seems to be the best suited to legislate or officially debate the effects and social implications of AI "what seems to characterise AI, to the point of introducing new elements into the debate on legal personality" they write "are the socio-technical profiles resulting from the deployment of artificial intelligence agents, e.g., the marked unpredictability of their decision-making processes and the impact (both positive and negative) that these processes may have on people's lives, society, and the market" (Novelli 2022). A legal personality status will clarify legal obligations for AI.

The second epistemic argument says that legal personhood gives AI entities a clear legal standing. The involvement of different human players in the production and implementation of such systems makes it difficult, if not impossible, to track the relevant player. It is, therefore, epistemically better to consider an AI system as a legal person, as this makes proper responsibility attributions possible.

These two arguments, importantly, seem to be the only ones to avoid the ontological discussion (a legal person does not need to be ontologically a person) or even the ontological comparison between existing legal persons like corporations and AI (the reason why legal personhood was granted to corporations does not require that the new entities to which it is granted resemble corporates in given ways). They also do not seek justification on the receiver's side of the communication by assuming, for instance, that naive users see AI a certain way.

Does this mean, then, that these two arguments can serve as a justification for using human terms like trustworthy to describe AI? A positive answer would need two qualifications: First, the texts using such expressions should explicitly stress that the sense of "trust" they used is done by analogy with the trust granted to legal persons and not actual persons. The person here is a legal bearer of responsibilities and rights, not a sentient human with all other properties. So far, most texts use the term without such explicit stress and even perhaps reference to legal persons. Second, even though the concept of a legal person remains legal fiction, it has real pragmatic and epistemic effects that continue to be debated. Attributions of trustworthiness to AI within such a framework can only be justified if the benefits of using such fiction outweigh the costs or risks—something which legal experts should assess more finely.

7 Conclusions

The language used to describe AI is under scrutiny. With terms like "trustworthy" being thrown around, it's no wonder there's confusion about what AI is and what it can do. But this loose talk comes with a risk. Could the wrong framing of AI lead to bad consequences or violates some basic principles of good institutional communication? Are institutions and experts justified in using and spreading these terms? These are important questions that need to be considered. After careful review, it seems there are few strong arguments for a human-leaning description of AI.

What about, one could object, the many fast or specific developments of AI that we are witnessing? There is a risk, already pointed out by Dreyfus in his famous paper on "Alchemy and Artificial Intelligence", that philosophers speak about AI and do not "define what sort of machines [they] have in mind….credulously assuming that highly intelligent artefacts have already been developed" (Dreyfus 1965, p. 2). The criticism partly holds here, as no specific AI system is described. Intelligence, however, is already one such human-leaning term, and what is being argued against extending the word "trustworthy" to AI here should perhaps be discussed already about the very use of the term "intelligence" in artificial intelligence.

Funding Open Access funding enabled and organized by Projekt DEAL. Ophelia Deroy is funded by the BIDT Grant Co-Sense and the Emerge EIC project (101070918).

Declarations

Conflict of interest The author has no conflict of interest to declare.

References

Bisio A, Sciutti A, Nori F, Metta G, Fadiga L, Sandini G, Pozzo T (2014) Motor contagion during human-human and human-robot interaction. PLoS ONE 9(8):e106172

Broadbent E (2017) Interactions with robots: the truths we reveal about ourselves. Ann Rev Psychol 68:627–652

Burton JW, Stein MK, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. J Behav Decis Mak 33(2):220–239

Chaminade T, Rosset D, Da Fonseca D, Nazarian B, Lutcher E, Cheng G, Deruelle C (2012) How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. Front Hum Neurosci 6:103

Coeckelbergh M (2011) Humans, animals, and robots: a phenomenological approach to human-robot relations. Int J Social Robot 3:197–204

Crawford K (2021) The atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, London

Dan-Cohen M (2016) Rights, persons, and organizations: a legal theory for bureaucratic society, vol 26. Quid Pro Books, London

De Graaf MM, Malle BF (2019, March) People'sex planations of robot behavior subtlyreveal mental state inferences. In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction(HRI). IEEE (pp. 239–248)

Dechêne A, Stahl C, Hansen J, Wänke M (2010) The truth about the truth: a meta-analytic review of the truth effect. Personal Soc Psychol Rev 14(2):238–257

Dehaene S, Lau H, Kouider S (2021) What is consciousness, and could machines have it?. Robotics, AI, and Humanity: Science, Ethics, and Policy, 43–56

Dennett DC (2019) What can we do. In: Brockman J (ed) Possible minds: twenty-five ways of looking at AI. Penguin Books, New York, p 41–53

Deroy O (2021) Rechtfertigende Wachsamkeit gegenüber KI. Künstliche Intelligenz—Die große Verheißung, vol 8. MoMo Berlin Philosophische KonTexte, Series, pp. 471–488

Dezecache G, Dockendorff M, Ferreiro DN, Deroy O, Bahrami B (2022) Democratic forecast: small groups predict the future better than individuals and crowds. J Exp Psychol: Appl 28(3):525–537

Dhami MK, Mandel DR (2021) Words or numbers? Communicating probability in intelligence analysis. Am Psychol 76(3):549

Dreyfus HL (1965) Alchemy and artificial intelligence. RAND Corporation, Santa Monica

Eyssel F, Hegel F (2012) (s)he's gott he look: Gender stereotyping of robots 1. J Appl Soc Psychol 42(9):2213–2230

Eyssel F, Kuchenbrandt D (2012) Social categorization of social robots: Anthropomorphism as a function of robot group membership. Br J Soc Psychol 51(4):724–731

Ferreiro D, Deroy O, Bahrami B (2023) Compromising improves forecasting. Royal Society Open Science

Floridi L (2019) Establishing the rules for building trustworthy AI. Nat Mach Intell 1(6):261–262

Geiselmann R, Tsourgianni A, Deroy O, Harris L (2023) Interacting with agents without a mind: the case for artificial agents. Curr Opin Behav Sci 49:101242

Goldberg S (2020) Trust and reliance. In: Simon J (ed) The Routledge Handbook of Trust and Philosophy. Routledge, New York, p 8

Gray HM, Gray K, Wegner DM (2007) Dimensions of mind perception. Science 315(5812):619–619

Hmamouche Y, Ochs M, Prévot L, Chaminade T (2020, October). Neuroscience to investigate social mechanisms involved in human-roboti nteractions. In Companion Publication of the 2020 International Conference on Multimodal Interaction. (pp. 52–56)

Hoffman MB (2014) The punisher's brain: the evolution of judge and jury. Cambridge University Press, Cambridge

Ishowo-Oloko F, Bonnefon JF, Soroye Z, Crandall J, Rahwan I, Rahwan T (2019) Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. Nat Mach Intell 1(11):517–521

John S (2018) Epistemic trust and the ethics of science communication: against transparency, openness, sincerity and honesty. Social Epistemology 32(2):75–87

Kahn P, Kanda T, Ishiguro H, Freier N, Severson R, Gill B, Ruckert J, Shen S (2012a) "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. Dev Psychol 48(2):303–314

Kahn P, Kanda T et al (2012b) Do people hold a humanoid robot morally accountable for the harmit causes? In Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (pp.33–40)

Kahn P, Shen S (2017) NOC NOC, who's there? A new ontological category (NOC) for social robots. In: Budwig N, Turiel E, Zelazo PD (eds) New perspectives on human development. Cambridge University Press, Cambridge, pp 106–120

Karpus J, Krüger A, Verba JT, Bahrami B, Deroy O (2021) Algorithm exploitation: humans are keen to exploit benevolent AI. iScience 24(6):102679

Keijsers M, Kazmi H, Eyssel F, Bartneck C (2021) Teaching robots a lesson: determinants of robot punishment. Int J Soc Robot 13:41–54

Kelley MS, Noah JA, Zhang X, Scassellati B, Hirsch J (2021) Comparison of human social brain activity during eye-contact with another human and a humanoid robot. Front Rob AI 7:599581

Kuchenbrandt D, Eyssel F, Bobinger S, Neufeld M (2013) When a robot's group membership matters: anthropomorphization of robots as a function of social categorization. Int J Soc Robot 5:409–417

Kuehne LM, Olden JD (2015) Lay summaries needed to enhance science communication. Proc Natl Acad Sci 112(12):3585–3586

Law T, Scheutz M (2021) Trust: recent concepts and evaluations in human-robot interaction. In: Scheutz M (ed) Trust in human-robot interaction. Springer International Publishing, New York, pp 27–57

Lee MK, Kiesler S, Forlizzi J, Srinivasa S, Rybski P (2010, March) Gracefully mitigating breakdowns in robotic services. In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp 203–210). IEEE

Lee M, Ruijten P, Frank L, de Kort Y, IJsselsteijn W (2021, May). People may punish, but not blame robots. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–11). Association for Computing Machinery

Medvecky F, Leach J (2019) An ethics of science communication. Springer Nature, NewYork

Novelli C (2022) Legal personhood for the integration of AI systems in the social context: a study hypothesis. AI & SOCIETY 1–19

Nyholm S (2023) Robotic animism: the Ethics of attributing minds and personality to Robots with Artificial Intelligence. In: Animism and Philosophy of Religion. Springer International Publishing, Cham, pp 313–340

Perez-Osorio J, Wykowska A (2020) Adopting the intentional stance toward natural and artificial agents. Philos Psychol 33(3):369–395

Perrig SA, Scharowski N, Brühlmann F (2023, April). Trust issues with trust scales: examining the psychometric quality of trust measures in the context of AI. In Extended abstracts of the 2023 CHI Conference on human factors in computing systems (pp. 1–7)

Popa E (2021) Human goals are constitutive of agency in artificial intelligence (AI). Philos Technol 34(4):1731–1750

Rauchbauer B, Nazarian B, Bourhis M, Ochs M, Prévot L, Chaminade T (2019) Brain activity during reciprocal social interaction investigated using conversational robots as control condition. Philos Trans Royal Soc B 374(1771):20180033

Salomons TV, Iannetti GD (2022) Fetal pain and its relevance to abortion policy. Nat Neurosci 25(6):879–881

Searle JR (1980) Minds, brains, and programs. Behav Brain Sci 3(3):417–424

Shank D, DeSanti A (2018) Attributions of morality and mind to artificial intelligence after real-world moral violations. Comput Hum Behav 86:401–411

Solaiman SM (2017) Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. Artif Intell Law 25:155–161

Sperber D (1982) Apparently irrational beliefs. In: Lukes S, Hollis M (eds) Rationality and relativism. Blackwell, Hoboken, pp 149–180

Sperber D (1997) Intuitive and reflective beliefs. Mind Lang 12(1):67–83

Sperber D, Norenzayan A, Shariff A, Gervais WM (2009) The cultural evolution of pro social religions. Behav Brain Sci 32(6):534–535

Tetlock PE, Gardner D (2015) Superforecasting: the art and science of prediction. Crown Publishers, New York

Thellman S, Silvervarg A, Ziemke T (2017) Folk-psychological interpretation of human vs. humanoid robot behavior: exploring the intentional stance toward robots. Front Psychol 8:1962

Tidoni E, Holle H, Scandola M, Schindler I, Hill L, Cross ES (2022) Human but not robotic gaze facilitates action prediction. Iscience 25(6):104462

Tomasello M (2022) The evolution of agency: behavioral organization from lizards to humans. MIT Press, Cambridge

Véliz C (2021) Moral zombies: why algorithms are not moral agents, vol 36. AI & society, pp. 487–497