AI and Ethics (2024) 4:691–698
https://doi.org/10.1007/s43681-024-00419-4

ORIGINAL RESEARCH

Anthropomorphism in AI: hype and fallacy

Adriana Placani[1]

Received: 27 October 2023 / Accepted: 4 January 2024 / Published online: 5 February 2024

Abstract
This essay focuses on anthropomorphism as both a form of hype and fallacy. As a form of hype, anthropomorphism is shown to exaggerate AI capabilities and performance by attributing human-like traits to systems that do not possess them. As a fallacy, anthropomorphism is shown to distort moral judgments about AI, such as those concerning its moral character and status, as well as judgments of responsibility and trust. By focusing on these two dimensions of anthropomorphism in AI, the essay highlights negative ethical consequences of the phenomenon in this field.

Keywords AI · Anthropomorphism · Ethics · Moral judgment · Fallacy · Hype

1 Introduction

The roots of anthropomorphism run deep. In the eighteenth century, David Hume wrote that there is a "universal tendency among mankind to conceive all beings like themselves and to transfer to every object, those qualities... and by a natural propensity, if not corrected by experience and reflection, ascribe malice or good-will to every thing, that hurts or pleases us" [1]. The long-standing phenomenon of anthropomorphism is still present today and one if its newest incarnations is in the field of artificial intelligence (AI).

There are many examples of anthropomorphism in the AI field, but perhaps the most famous instantiation of it is the "ELIZA effect". ELIZA, considered the first chat bot, was a natural language processing program developed by Joseph Weizenbaum at MIT in the 1970s. In spite of the unusually constrained form of dialogue used by ELIZA [2], which consisted of simply mirroring or rearranging whatever a user said in the style of a Rogerian psychotherapist, people related to the program in anthropomorphic ways as though it was a person [3]. As Weizenbaum wrote: "What I had not realized is that extremely short exposure to a relatively simple computer program could induce powerful delusional thinking in quite normal people" [3]. Subsequently, Weizenbaum spent much of his life warning about the dangers of projecting human qualities onto AI.

In a similar vein, this essay offers an examination of anthropomorphism in AI by focusing primarily on some of its negative ethical consequences. An exhaustive analysis of such consequences would be virtually impossible, but by focusing on anthropomorphism construed as a form of hype and as a fallacy this work shows that and how anthropomorphism overinflates the capabilities and performance of AI systems, as well as distorts a host of moral judgments about them.

This work is structured as follows. In the first section, the paper explains what anthropomorphism entails, as well as some of ways in which the phenomenon manifests itself in the field of AI. Emphasis is placed here on showing that anthropomorphism is a constitutive part of the hype surrounding AI. Hype in this context is understood as the misrepresentation and over-inflation of AI capabilities and performance, while being a constitutive part of hype is understood as being a part of the creation of hype. In the second section, the essay shows that anthropomorphism distorts moral judgments through its fallacious character. It illustrates this by focusing on four central moral judgments about AI: judgments concerning its moral character and status, as well as judgments about responsibility and trust in AI. The third section ends this work by providing a brief summary and conclusion.

✉ Adriana Placani
adrianaplacani@fcsh.unl.pt

[1] Institute of Philosophy (IFILNOVA), Nova University of Lisbon, Lisbon, Portugal

2 Anthropomorphism and hype about AI

This section briefly describes anthropomorphism in general terms, only to focus in more detail on its manifestations in the AI field. The aim here is to show that and how anthropomorphism can be construed as a form of hype in virtue of it misrepresenting, distorting and exaggerating AI capabilities and performance.

Anthropomorphism is the ascription of human qualities (e.g., intentions, motivations, human feelings, behaviors) onto non-human entities (e.g., objects, animals, natural events) [4, 5]. This phenomenon is considered an evolutionary and cognitive adaptive trait [6], which does not necessarily correlate to the features of that which is anthropomorphized [4]. Instead, it represents a distinctively human process of inference or interpretation [7] that includes not only perceiving an entity as human-like in terms of its physical features, but also imbuing it with mental capacities considered uniquely human, such as emotions (e.g. empathy, revenge, shame, and guilt) and the capacity for conscious awareness, metacognition and intention formation [8].

Anthropomorphism is a pervasive and widespread phenomenon that garners new dimensions in the realm of AI. One dimension that is seldom emphasized relates to the hype surrounding AI systems. In virtue of the attribution of distinct human characteristics that misrepresent and exaggerate AI capabilities and performance, anthropomorphism in AI can be viewed as a constitutive part of hype. To see this, consider first anthropomorphic language.

Anthropomorphic language is so prevalent in the discipline that it seems inescapable. Perhaps part of the reason is because anthropomorphism is built, analytically, into the very concept of AI. The name of the field alone—artificial intelligence—conjures expectations by attributing a human characteristic—intelligence—to a non-living, non-human entity, which thereby exposes underlying assumptions about the capabilities of AI systems. Using such anthropomorphic language also invites interpreting algorithmic behavior as human-like so that it may be compared to human modes of reasoning [9].

Going beyond the concept, there are many examples of anthropomorphic language that exaggerate the capabilities of AI, starting from the earliest days of the field. Alan Turing, creator of the Turing test, among other things, described his machines in anthropomorphic terms in spite of the fact that they were simple abstract computational devices. For example, he compared what he dubbed his 'child-machine' to Helen Keller and said that the machine could not 'be sent to school without the other children making excessive fun of it', but that it would get 'homework' [10].

Famed cyberneticist Valentino Braitenberg also used anthropomorphisms to describe his very simple robot vehicles, which were said to dream, sleepwalk, have free will, 'ponder over their decisions', be 'inquisitive', 'optimistic', and 'friendly' [10]. Other researchers, such as David Hogg, Fred Martin, and Mitchel Resnick used anthropomorphic language for their robots even though these robots were built from LEGO bricks containing electronic circuits. Masaki Yamamoto described his vacuum cleaner robot, Sozzy, as 'friendly' and as having 'four emotions... joy, desperation, fatigue, sadness' [10].

More recently, Sophia, a robot with a human-like form was granted citizenship in Saudi Arabia, was a guest on various TV shows and news programs, and appeared beside world leaders and policymakers. As Sven Nyholm [11–13]. Writing about Sophia, computer scientist Noel Sharkey highlighted that "it is vitally important that our governments and policymakers are strongly grounded in the reality of AI at this time and are not misled by hype, speculation, and fantasy" [13].

The examples above show how anthropomorphisms have been part and parcel of the hype surrounding AI in robotics and, indeed, anthropomorphism is a well-known and well-researched phenomenon in this area. After all, human characteristics are used as guiding principles in robot design, while perceiving robots as humanlike is important to human–robot interactions [14, 15]. However, this should not lead to the conclusion that anthropomorphism in the AI field is isolated to robotics.

Anthropomorphism has also been displayed around deep neural networks (DNNs). In 2022, Ilya Sutskever, co-founder and chief scientist at OpenAI, hyped up DNNs by declaring: "it may be that today's large neural networks are slightly conscious" [16]. It is true that DNNs are one of the most advanced and promising fields within AI research, with DNN architecture applied in AlphaZero's famous win over the human Go world champion and a part of many AI-related applications, such as Google translation services, Facebook facial recognition software, and virtual assistants like Apple's Siri [9, 17]. However, in spite of the many accomplishments achieved using deep neural networks, parallels to the human brain should be resisted.

Shimon Ullman [18] argues that almost everything we know about neurons (e.g., structure, types, interconnectivity) has *not* been incorporated in these networks [17, 18]. DNNs use a limited set of highly simplified homogeneous artificial "neurons", whereas biological neuronal architecture displays a heterogeneity of morphologies and functional connections [17, 18]. Thus, describing network units in anthropomorphic terms as, for example, biological neurons is an enormous simplification given the highly sophisticated nature and diversity of neurons in the brain [17, 19]. Conversely, it is also an over-inflation of DNNs' capabilities.

The New York Times' 2018 article on AlphaZero's victories is a good example of anthropomorphic tendencies that seem to do just that—overinflate capabilities:

> Most unnerving was that AlphaZero seemed to express insight. It played like no computer ever has, intuitively and beautifully, with a romantic, attacking style. It played gambits and took risks. In some games it paralyzed Stockfish [the reigning computer world champion of chess] and toyed with it . . . AlphaZero had the finesse of a virtuoso and the power of a machine. It was humankind's first glimpse of an awesome new kind of intelligence. . . AlphaZero won by thinking smarter, not faster . . . It was wiser, knowing what to think about and what to ignore [20].

Part of the problem with anthropomorphic language as exhibited above is that it asserts an out-of-place human-centric perspective that conceals the reality of how these networks work, as well as their limitations. David Watson [9], for example, has argued that DNNs' similarities to human cognition have been seriously overstated and narrowly construed, especially in light of DNNs' considerable shortcomings (e.g., brittleness, inefficiency, and myopia).

A final example of anthropomorphism that exaggerates AI capabilities and performance comes from large language models (LLMs). LLMs, such as ChatGPT, Bing Chat (Sydney) and LaMDA have garnered a lot of attention recently. These AI-powered chat bots belong to the class of AI called generative AI, are trained on vast amounts of data, use artificial neural networks and can generate human-like responses to any question users can think of. Given the latter, it almost seems like hype through anthropomorphism was bound to happen.

For example, in a recent cross-sectional study of 195 randomly drawn patient questions, a team of licensed health care professionals compared physicians and ChatGPT's responses to patients' questions [21]. The chat bot responses were preferred over physician responses and rated significantly higher for both quality and empathy [21]. Importantly, the proportion of responses rated empathetic or very empathetic was significantly higher for the chat bot than for physicians, amounting to a 9.8 times higher prevalence of empathetic or very empathetic responses for the chat bot [21]. This means that almost half of responses from ChatGPT were considered to be empathetic (45%) compared to less than 5% of those from physicians.

This example is noteworthy because attributing empathy to a chat bot anthropomorphizes the latter since empathizing is a complex emotional and cognitive process that involves the ability to recognize, comprehend and share the feelings of others.

Other notorious examples of anthropomorphizing chat bots include the infamous exchange between Sydney, Microsoft's chatbot, and the New York Times' technology columnist Kevin Roose [22] and the declaration by Blake Lemoine, a Google engineer, that the company's chat bot, LaMDA, was conscious and capable of feelings [23].

Given that AI is far from being sentient now, anthropomorphisms such as these fan the flames of hype by misrepresenting the current state of AI systems and potentially leading to mistaken beliefs, as well as overblown fears and hopes. AI already exhibits a great deal of influence in our world, and this is only going to continue to grow. Exaggerating the capabilities of these systems conceals the reality of AI achievements and impedes their understanding. This leads to a generalized lack of knowledge about how these systems work, which can feed extreme beliefs and sentiments through misinformation. On the other hand, the phenomenon is also reductive because it asserts an out-of-place, bio-centric perspective that can overlook the unique potential of artificial systems.

Projecting human capabilities onto artificial systems is a relatively new manifestation of a long-standing and natural phenomenon, but in the realm of AI, this may lead to serious ramifications. The above offers some telling examples of anthropomorphism in AI but does not and indeed cannot provide an exhaustive account of this phenomenon or its connection to hype. Nevertheless, it seems fair to conclude that anthropomorphism is part of the hype surrounding AI systems because of its role in exaggerating and misrepresenting AI capabilities and performance. Furthermore, such over-inflation and misrepresentation is nothing mysterious. It is simply due to projecting human characteristics onto systems that do not possess them.

3 Anthropomorphism and moral judgments about AI

The previous section showed the prevalence of anthropomorphism as exhibited across the field by researchers, developers, science communicators, and the public. It also showed that the pervasiveness of this phenomenon is nothing new. However, in spite of the fact that anthropomorphism is a well-known occurrence, its ethical consequences are less understood.

Anthropomorphism is also a kind of fallacy, and this is often overlooked. The fallacy occurs when one assumes or makes the unwarranted inference that a non-human entity has a human quality. This can involve projecting human characteristics onto non-humans, such as: "My car is angry at me" or making an unwarranted inference about non-humans, such as "The robot is friendly because it waved at me". In this way, anthropomorphism can be regarded as either a factual error—when it involves the attribution of a human characteristic to some entity that does not possess that characteristic,

or as an inferential error—when it involves an inference that something is or is not the case when there is insufficient evidence to draw such a conclusion [24].

As a kind of fallacy, then, anthropomorphism involves a factually erroneous or unwarranted attribution of human characteristics to non-humans.[1] Given this, when anthropomorphism becomes part of reasoning it leads to unsupported conclusions. The following will discuss some of these conclusions and how they occur within moral judgment. In this way, some of the negative ethical implications of anthropomorphizing AI will be exposed.

There is a necessary connection between attributing human traits to AI and a distorting effect on various moral judgments about AI. This distorting effect occurs because attributing human characteristics to AI is currently fallacious, affecting beliefs and attitudes about AI, which in turn play a role in moral judgment.[2] The activity of moral judgment is that of reasoning, deliberating, thinking about whether something has a moral attribute [25]. The thing assessed might be an action, person, institution or state of affairs, and the attribute might either be general (such as rightness or badness) or specific (such as loyalty or injustice) [25].

For example, consider how anthropomorphic language (e.g., AI systems "learn", decision algorithms "think", classification algorithms "recognize", Siri and Alexa are "listening") can influence deliberation, be it moral or otherwise. Such language shapes how we think about AI because it provides us with the conceptual framework, tools, and terminology for forming, expressing and organizing our beliefs, expectations and general understanding of these systems. In other words, it is how we conceptualize AI.

Furthermore, anthropomorphic language stands to influence both conscious and unconscious thinking about AI. Although it might be thought that only conscious reflection plays a central role in moral judgments, Haidt [26], for example, has argued that quick, automatic processes drive moral judgment while reflective processes play a more ad hoc role. Consequently, even if, on reflection, one might actively avoid anthropomorphic language when engaging in moral reasoning, it is possible that moral judgments are still distorted by it.

Furthermore, consider perhaps the biggest problem with anthropomorphizing AI, which is that viewing AI as human-like involves viewing it as having human-like agency. To be clear, at this point at least, conceiving of AI as having this kind of agency is a mistake because human agency involves having the capacity to act intentionally, where intentional actions are explained in terms of mental states (e.g., beliefs, desires, attitudes) that are the causal antecedents of an agent's behavior [27]. No such mental states could be attributed, plausibly, to current AI, which means that attributing this kind of agency to AI systems is a mistake and not an isolated one.[3] This error can have serious consequences because it can distort moral judgments about AI. When an error such as this becomes part of moral reasoning, then arguments based (or partly based) on it become fallacious and any subsequent conclusion unfounded.

To appreciate the distorting effects of anthropomorphism, the following will consider four moral judgments about AI systems: judgments of moral character, judgments of moral status, responsibility judgments and judgments of trust. To be clear, these moral judgments are distorted not necessarily in their verdict, but in the process of arriving at their verdict when this process is (partly) based on anthropomorphism. Furthermore, these moral judgments are to be addressed in turn even though there are many points of convergence between them. Finally, it should be noted that a full treatment of such extensive moral issues is not possible given their breadth, but that, nevertheless, the following seeks to illustrate how anthropomorphism affects them in virtue of the attribution of human-level qualities onto entities that do not possess them.

3.1 Judgments of moral character

Dating back to Aristotle, moral character is, primarily, a function of having or lacking various virtues and vices. The virtues and vices that comprise one's moral character are typically understood as dispositions to behave in certain ways in certain sorts of circumstances [28]. Thus, a moral character judgment can be defined as an evaluation of another's moral qualities, i.e., their virtues and vices.

Making moral character judgments about other people is a common practice. The way in which such judgments are made differs, but it typically involves, at least, three sources of information: about another's behavior, their perceived mind and their identity [29]. Thus, character judgments hinge on what others do, what they seem to think and on

[1] This does not deny that it is possible for non-humans to possess human characteristics. However, as a fallacy, anthropomorphism necessarily involves a kind of an error. Indeed, charges of anthropomorphism usually imply some kind of mistaken attribution of human traits [24].

[2] Whether AI could develop human qualities (e.g., awareness) is an open question.

[3] There are scholars [44–46] who suggest that the concepts of agency and moral agency should be broadened and intentions not taken into account. They argue in favor of artificial or virtual agency and even artificial or virtual moral agency, but they do not claim that these kinds of agencies are human-like. For example, Floridi [47] claims that to be called 'agents' systems have to be interactive, autonomous, and adaptive and that all 'agents' whose actions have morally qualifiable consequences are 'moral agents' [46]. For a useful criticism of this view, see Fritz, et al. [48].

who these others are (e.g., in terms of appearance, group membership) [29].

Normally, the second criteria for making character judgments—the perceived mind of others—disqualifies non-human entities from being the subject of moral character judgments [29]. This is because when making character judgments about others, one must make inferences about their minds, which includes making inferences about their intentions and moral capacities [29]. In the case of AI systems, the absence of mental states, their inability to understand moral issues or reason about morality should disqualify them from being the subjects of such judgments.

However, anthropomorphism can change all that. In fact, the previous section offered some examples of moral character judgments of AI, such as the 'empathetic' ChatGPT, the 'friendly' Sozzy robot, and the 'wise' AlphaZero. This means not only that anthropomorphism can distort moral judgments, but also that it can distort them to such an extent that a previously inappropriate evaluation becomes appropriate. By projecting a mind onto AI systems, AI becomes the subject of moral character evaluations.

If AIs are perceived as having mental states, then they can be characterized in moral terms as good, evil, friendly, empathetic, wise, loyal, courageous, bad, trustworthy, etc. In other words, the whole plethora of virtues and vices, which are said to make up moral character, becomes available. In the absence of moral agency, this is problematic. For example, on Aristotle's view, a virtuous agent is not one that just performs virtuous actions, but also one that understands those actions, whose actions result from a fixed character, and who chooses the action in question "for its own sake" (e.g., the agent chooses to be honest because they believe there is something intrinsically good about being honest) [30, 31]. These criteria are far from the capabilities of current AI systems, which means that attributing virtues to them is troublesome.

Moreover, trouble compounds because character is usually perceived as a partial driver of future moral behavior. For example, a person judged to be 'evil' will probably be perceived as more likely to do evil things, while a person judged to be 'good' will probably be perceived as more likely to do good things [29]. This means that AIs perceived in moral terms will also be perceived as more or less likely to behave in accordance with their so-called virtues and vices. This, in turn, can affect human interactions with AI systems, as well as human dispositions, expectations and attitudes towards AI (e.g., of trust, hope, suspicion). Needless to say, these would be as supported as the attribution of virtues and vices on the basis of anthropomorphism. That is, not at all.

3.2 Judgments of moral status

An entity with moral status is one that matters (to some degree) morally in and of itself [32]. More precisely, if an entity has moral status, then there are certain moral reasons or requirements concerning how it is to be treated for its own sake [32]. Thus, to have a moral status is to be an entity towards which moral agents have, or can have, moral obligations [33].

Arguably, the moral status of an entity should be based on the intrinsic properties of that entity [34].[4] In the 2020 book, *Ethics of Artificial Intelligence*, Matthew Liao [34] provides the following list of empirical, non-speciesist, intrinsic properties that could afford moral status to an entity: being alive; being conscious; being able to feel pain; being able to desire; being capable of rational agency (e.g., being able to know something about causality, such as if one does x, then y would happen, and being able to bring about something intentionally); being capable of moral agency, such as being able to understand and act in light of moral reasons. If some or all of these characteristics are present in an entity, then moral status could be afforded to it.

It is clear that anthropomorphizing AIs can involve the projection of some of these qualities onto AI systems. Anthropomorphism is the attribution of intrinsic human qualities onto non-humans. The problem is that these qualities can then become a part of moral judgment. For example, viewing AI as human-like could project consciousness onto AI systems. We have seen that this actually happens in the previous section. However, viewing AI as conscious is not contained to just this, but can become a reason in favor of conferring moral status onto it.

Anthropomorphism raises problems for the kinds of evidence we need to make inferences about the moral status of entities like AI. John Danaher [35], a proponent of ethical behaviorism, claims that a sufficient ground for believing that an entity has moral status is that it is roughly behaviorally equivalent to another entity of whose moral status we are already convinced (e.g., humans).[5] However, judgments of behavioral equivalency can easily be undermined by the tendency to project human qualities onto AI.[6]

The projection of human traits onto AI systems can bestow AI with moral status when it is not deserved through the attribution of intrinsic qualities that are not present in

[4] Cf. Mark Coeckelbergh [49] who advances a relational approach to moral status, which affords the latter based on social relations between different entities (e.g., human beings and robots).

[5] Cf. Shevlin [50] who argues in favor of cognitive equivalence, which is the view that we ought to regard AI as a psychological moral patient if it is cognitively equivalent to beings we already regard as psychological moral patients.

[6] Thank you to an anonymous reviewer for suggesting this implication.

AI, but that are criteria for moral status (e.g., capacity for rational agency, capacity for feeling). Granting moral status to AIs would then make them into a site of moral concern, as well as make them into potential right-holders to whom duties are owed. Of course, if AIs come to possess some or all of the intrinsic qualities mentioned, then it is plausible to afford them the same kind of moral status as other entities that have such properties [34]. Until then, however, such judgments are flawed.

3.3 Responsibility judgments

Another moral judgment that can be affected by anthropomorphism concerns attributions of moral responsibility. If AIs are attributed certain human capacities, then their possession could qualify them as morally responsible agents. For example, if AI is perceived as having a mind of its own, then this means that it can be viewed as capable of intentional action, and therefore, held responsible for its actions [36]. For an entity to be morally responsible for its actions, it has to be a moral agent. At this point, it is important to note that having moral status, which was the judgment considered before, and being a moral agent are distinct. An entity is a moral agent when it is morally responsible for what it does. For example, a baby is not a moral agent because it lacks moral competency, but it does have a moral status as it is considered a moral patient that can be wronged.

If AI is perceived as a moral agent, then it can be held responsible and blamed or praised for its actions, as well as for the consequences of its actions. However, blaming or praising an AI would be futile given the absence of any kind of understanding of such moral responses on the part of AI systems. The error is in regarding the AI itself as a site of moral responsibility in the absence of moral agency. Moral agency requires that one can meet the demands of morality. This requirement is interpreted in different ways on different accounts: as being able to obey moral laws, act for the sake of the moral law, have an enduring self with free will and an inner life, understand relevant facts as well as have moral understanding, have a capacity for remorse and concern for others [25]. Arguably, until AI can meet such requirements, they should not be considered moral agents.[7] Through the process of anthropomorphization, however, AIs can be attributed qualities that render a verdict of moral agency plausible when it is not.

One of the dangers of anthropomorphizing AI in this way is that judgments about responsibility might then focus solely on AI as a locus of responsibility. This is a problem in and of itself because of the error in agential attribution on anthropomorphic bases, but also because it has the potential to absolve others of responsibilities that they ought to bear by shifting focus onto AIs themselves. For example, holding an AI to be responsible for some action or outcome can obfuscate the responsibilities of potentially rightful bearers of responsibility, such as those of owners, developers or governments who each have distinct roles to play in, respectively, responsible ownership, development and regulation.[8]

If AIs are viewed as having a mind of their own, then this can lead not only to a distorted judgment of responsibility that sees the AIs as responsible, but also to others being wrongfully absolved of responsibility, as well as to a generalized sense of a loss or outright lack of control. The latter effect is because if AIs are moral agents, then they are also autonomous decision-makers who are able to choose their own goals and act freely in light of moral reasons. This would mean that their decisions are outside of human control, as well as (usually) opaque because of the black-box nature of AI algorithms, but that they ought to be given the same moral weight and respect as the decisions of any other moral agent's. However, until AIs become such agents (if ever), such moral judgments are flawed.

3.4 Judgments of trust

Trust is an attitude towards those (or that) which we hope is trustworthy, where trustworthiness is a property not an attitude [37]. Trust and trustworthiness are distinct, but, ideally, what is trusted is trustworthy, and what is trustworthy is trusted [37]. However, it is clear that trust can be misplaced. At a very basic level, trust is about a trustor that trusts (judges the trustworthiness of) a trustee with regard to some object of trust [38]. Trustworthy trustees are those worthy of being trusted. To be worthy of trust, they must be capable of being trusted, which means that they must have the competence to fulfil the trust that is placed in them [39].

One of the dangers of anthropomorphizing AI is that judgments about whether to trust AI can become judgments concerning the trustworthiness of the AI itself. However, according to two of the most prevalent conceptions of trust, AIs are not capable of being trusted [39]. On affective accounts of trust, 'trust is composed of two elements: an affective attitude of confidence about the goodwill and

[7] Cf. Floridi [47] who argues that all 'agents' whose actions have morally qualifiable consequences are 'moral agents', while to be called 'agents' systems have to be interactive, autonomous, and adaptive. According to Floridi, such 'moral agents' without intentions are not morally responsible, but they are accountable (e.g., they can be modified, deleted, disconnected) [46, 48]. Fritz, et al. [48] criticize this view of 'moral agency' without moral responsibility as an empty concept.

[8] Cf. Bryson, Diamantis, and Grant [51] who argue in a similar vein, but about the dangers of granting AI legal personhood, that natural persons could use artificial persons to shield themselves from the consequences of their conduct and Rubel, Castro, and Pham [52] who argue that enlisting technological systems into agents' decision-making processes can obscure moral responsibility for the results.

competence of another... and, further, an expectation that the one trusted will be directly and favorably moved by the thought that you are counting on them' [40]. However, AI lacks the capacity to be moved by trust or a sense of goodwill since it lacks any emotive states [39]. On normative accounts, trustees need to be appropriate subjects of blame in those situations when trust is breached [39]. This means that trustees need to be able to understand and act on what is entrusted to them, as well as be held responsible for those actions [39]. However, AI is not a moral agent in any such standard sense, so it cannot he held morally responsible for its actions. Thus, AI lacks the capacity for being trusted on both of these accounts.

The problem then is that anthropomorphizing AIs can lead to viewing such systems as trustworthy. Certain qualities can be projected onto AI, such as goodwill, empathy, or moral agency, and on their bases, the wrong conclusion can be drawn. Conceiving of AIs themselves as trustworthy is, by itself, erroneous when based on such projected qualities, but it can also have additional effects. For example, a verdict that AI systems are trustworthy can obfuscate the degree of trustworthiness of other parties. This is because trusting AIs because AIs themselves are trustworthy can leave out of considerations factors that ought to include when making such verdicts, such as the trustworthiness of owners, of developers, of organizations behind the deployment of AIs or the trustworthiness of governments whose responsibility it is to regulate the industry. This means that regarding AIs as trustworthy is not only a problematic moral judgment, but also an obfuscation of important considerations that should factor in judgments of trust.

It should be noted that, in the literature, the idea that anthropomorphism in AI affects trust is present by way of empirical findings that support the view that anthropomorphism increases trust in AI. For example, in the context of both autonomous vehicles and virtual agents, people showed more trust in AIs with human characteristics than without [41–43]. In general, the claim is that the more human-like an AI agent is, the more likely humans are to trust and accept it [41]. If this is accurate, then this carries with it serious ethical implications as well because of the possibility of exploiting this human bias for manipulative or deceptive purposes.

4 Conclusion

This work has focused on anthropomorphism as a form of hype and as a fallacy. The first section showed how anthropomorphism tends to exaggerate and misrepresent AI capabilities by attributing human-like attributes onto systems that do not possess them. The second section showed that, via the same mechanism, anthropomorphism distorts moral judgments about AI, such as those concerning AI's moral character and status, as well as judgments of responsibility and trust in AI. In these ways, this work has shown some of the more acute negative consequences of anthropomorphism.

Funding Open access funding provided by FCT|FCCN (b-on). Adriana Placani's work is financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., under the Scientific Employment Stimulus - Individual Call - CEECIND/02135/2021.

Declarations

Conflict of interest The author has no competing interests to declare that are relevant to the content of this article.

References

1. Hume, D.: The natural history of religion. Stanford University Press, Stanford (1957)
2. Weizenbaum, J.: How does one insult a machine? Science **176**, 609–614 (1972)
3. Weizenbaum, J.: Computer power and human reason, from judgment to calculation. WH. Freeman, San Francisco (1976)
4. Airenti, G.: The cognitive basis of anthropomorphism: From relatedness to empathy. Int. J. Soc. Robot. **7**(1), 117–127 (2015)
5. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: a three-factor theory of anthropomorphism. Psychol. Rev. **114**(4), 864–886 (2007)
6. Ellis, B., Bjorklund, D.: Origins of the social mind: evolutionary psychology and child development. The Guildford Press, New York (2004)
7. Epley, N., Waytz, A., Akalis, S., Cacioppo, J.T.: When we need a human: motivational determinants of anthropomorphism. Soc. Cogn. **26**(2), 143–155 (2008)
8. Johnson, J.: Finding AI faces in the moon and armies in the clouds: anthropomorphising artificial intelligence in military human–machine interactions. Glob. Soc. **38**, 1–16 (2023)
9. Watson, D.: The rhetoric and reality of anthropomorphism in artificial intelligence. Mind. Mach. **29**, 417–440 (2019)
10. Proudfoot, D.: Anthropomorphism and AI: Turing's much misunderstood imitation game. Artif. Intell. **175**(5–6), 950–957 (2011)
11. Nyholm, S.: Humans and robots: ethics, agency, and anthropomorphism. Rowman & Littlefield International (2020)
12. Halpern, S.: The New Yorker," 26 July 2023. [Online]. https://www.newyorker.com/tech/annals-of-technology/a-new-generation-of-robots-seems-increasingly-human. Accessed 16 Oct 2023
13. Sharkey, N.: Mama Mia, It's Sophia: A Show Robot or Dangerous Platform to Mislead? Forbes, 17 November 2018. [Online]. https://

www.forbes.com/sites/noelsharkey/2018/11/17/mama-mia-its-sophia-a-show-robot-or-dangerous-platform-to-mislead/. Accessed 19 Oct 2023
14. Fink, J.: "Anthropomorphism and human likeness in the design of robots and human–robot interaction." In: Social robotics. 4th International Conference, ICSR 2012, Chengdu, (2012)
15. Rinaldo, K., Jochen, P.: Anthropomorphism in human–robot interactions: a multidimensional conceptualization. Commun. Theory **33**(1), 42–52 (2023)
16. Sutskever, I.: It may be that today's large neural networks are slightly conscious. Twitter, 9 February 2022. [Online]. https://twitter.com/ilyasut/status/1491554478243258368. Accessed 19 Oct 2023
17. Salles, A., Evers, K., Farisco, M.: Anthropomorphism in AI. A JOB Neurosci **11**(2), 88–95 (2020)
18. Ullman, S.: Using neuroscience to develop artificial intelligence. Science **363**(6428), 692–693 (2019)
19. Geirhos, R., Janssen, D., Schütt, H., Rauber, J., Bethge, M.: Comparing deep neural networks against humans: object recognition when the signal gets weaker, arXiv preprint https://arXiv.org/1706.06969, (2017)
20. Strogatz, S.: One giant step for a chess-playing machine, The New York Times , 26 12 2018. [Online]. https://www.nytimes.com/2018/12/26/science/chess-artificial-intelligence.html. Accessed 11 Oct 2023
21. Ayers, J., Poliak, A., Dredze, M., Leas, E., Zechariah, Z., Kelley, J., Dennis, F., Aaron, G., Christopher, L., Michael, H., Davey, S.: Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern. Med. **183**(6), 589–596 (2023)
22. Roose, K.: A conversation with Bing's Chatbot left me deeply unsettled. The New York Times, 16 February 2023. [Online]. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html. Accessed 20 Oct 2023
23. Tiku, N.: The Google engineer who thinks the company's AI has come to life. The Washington Post, 11 June 2022. [Online]. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/. Accessed 20 Oct 2023
24. Mitchell, R.W., Thompson, N.S., Miles, L.H.: Anthropomorphism, anecdotes, and animals. SUNY Press (1997)
25. Craig, E.: The shorter Routledge encyclopedia of philosophy. Routledge, New York (2005)
26. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol. Rev. **108**, 814–834 (2001)
27. Davidson, D.: The essential Davidson. Oxford University Press, New York (2006)
28. Timpe, K.: Moral character, Internet Encyclopedia of Philosophy (2007)
29. Hartman, R., Will, B., Kurt, G.: Deconstructing moral character judgments. Curr. Opin. Psychol. **43**, 205–212 (2022)
30. Milliken, J.: Aristotle's aesthetic ethics. South. J. Philos. **44**(2), 319–339 (2006)
31. Kelly, J.: Virtue and pleasure. Mind **82**(327), 401–408 (1973)
32. Jaworska, A., Julie, T.: The grounds of moral status. The Stanford encyclopedia of philosophy (2023)
33. Warren, M.: Moral status: obligations to persons and other living things. Clarendon Press, Oxford (1997)
34. Liao, S.M.: The moral status and rights of AI. In: Liao, S.M. (ed.) Ethics of artificial intelligence, pp. 480–505. Oxford University Press, Oxford (2020)
35. Danaher, J.: What matters for moral status: behavioral or cognitive equivalence? Camb. Q. Healthc. Ethics **30**(3), 472–478 (2021)
36. Waytz, A., Cacioppo, J., Epley, N.: Who sees human?: The stability and importance of individual differences in anthropomorphism. Perspect. Psychol. Sci. **5**(3), 219–232 (2010)
37. McLeod, C.: Trust. The Stanford encyclopedia of philosophy (2023)
38. Bauer, P.: Clearing the jungle: conceptualising trust and trustworthiness. In: Barradas-de-Freitas, R.A.S.L.I. (ed.) Trust matters: cross-disciplinary essays, pp. 17–34. Bloomsbury Publishing, Oxford (2021)
39. Ryan, M.: In AI we trust: ethics, artificial intelligence, and reliability. Sci. Eng. Ethics **26**, 2749–2767 (2020)
40. Jones, K.: Trust as an affective attitude. Ethics **107**(1), 4–25 (1996)
41. Waytz, A., Joy, H., Nicholas, E.: The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. J. Exp. Soc. Psychol. **52**, 113–117 (2014)
42. Kim, K., Boelling, L., Haesler, S., Bailenson, J.: Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR. In: IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Munich, (2018)
43. Verberne, F.M.F., Jaap, H., Cees, J.H.: Trusting a virtual driver that looks, acts, and thinks like you. Hum. Factors **57**(5), 895–909 (2015)
44. Coeckelbergh, M.: Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. AI & Soc. **24**, 181–189 (2009)
45. Floridi, L.: Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **374**(2083), 20160112 (2016)
46. Floridi, L., Sanders, J.: On the morality of artificial agents. Mind. Mach. **14**, 349–379 (2004)
47. Floridi, L.: Levels of ABSTRACTION AND THE TURING TEST. Kybernetes **39**, 423–440 (2010)
48. Fritz, A., Brandt, W., Gimpel, H., Bayer, S.: Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI). De Ethica **6**(1), 3–22 (2020)
49. Coeckelbergh, M.: The moral standing of machines: towards a relational and non-Cartesian moral hermeneutics. Philos. Technol. **27**(1), 61–77 (2014)
50. Shevlin, H.: How could we know when a robot was a moral patient? Camb. Q. Healthc. Ethics **30**(3), 459–471 (2021)
51. Bryson, J., Diamantis, M., Grant, T.: Of, for, and by the people: the legal lacuna of synthetic persons. Artif. Intell. Law **25**, 273–291 (2017)
52. Rubel, A., Castro, C., Pham, A.: Agency laundering and information technologies. Ethical Theory Moral Pract **22**, 1017–1041 (2019)