Anthropomorphism and AI Hype

-

Nicholas Barrow

ABSTRACT

As humans, we have an innate tendency to ascribe human-like qualities to non-human entities. Whilst sometimes helpful, such anthropomorphic projections are often misleading. This commentary considers how anthropomorphising AI contributes to its misrepresentation and hype. First, I outline three manifestations (terminology; imagery; and morality). Then, I consider the extent to which we ought to mitigate it.

Introduction

Humans have an innate tendency to anthropomorphise. We can't help but think about non-humans in distinctively human ways, 'seeing' human-like traits in non-humans, and responding accordingly [1]. But these inferences can be misleading. It is comforting and perhaps natural to infer that our pet is happy to see us when we come home. Such an inference might seem plausible for our pet dog. But less so for our pet rock.

With the advent of sophisticated AI, the anthropomorphism of technological artefacts has become widespread [2, 3]. This is not least because such technologies are often *designed* to be human-like. Indeed, in many cases, this is what propels their success. The popular LLM-based conversation app Replika, for instance, gained popularity *because* it felt to users like they were talking with a human. Replika users perceive their *Replika* to have certain human attributes like the capacity for emotion (for example, love, lust, happiness, and sadness). This is in part because a human-like avatar accompanies the chatbot. But it is also because of its ability to maintain human-level conversations and to do so using first-person pronouns, something Ben Schneiderman and Micheal Muller have critiqued GPT-4 about [4]. Social robots are designed to function within social situations and facilitate human-robot interaction. To this end, they are often modelled to look and behave like humans (and some higher-order primates like dogs) to *evoke* anthropomorphised responses [5].

Some express concern that taking advantage of anthropomorphisation in this way is deceptive [6]. Your *Replika* does not *really* have the capacity for emotion. Nor do social robots that behave and look like they have certain human traits *have* those traits.

Designing them in ways that make us think they do is deceptive. Whilst true, deception can serve important purposes: many users, for instance, rely on Replika for friendship, therapy, and intimacy; sometimes to combat loneliness. But this reliance has created a troubling power balance. In February 2023, *Luka* (the company behind Replika) decided to turn Replika's capacity for intimacy off. This left many users to experience real heartbreak over an artificial relationship [7].

To take a more infamous example: recall in June 2022 when Blake Lamoine (a then software engineer at Google) declared that LaMDA (a chatbot) was sentient [8]. Lamoine's evidence for this claim was based on testimony. They interviewed LaMDA who revealed it was aware of its existence, was a person, and asked not to be turned off. At one point, Lamoine expresses concern: "I could be wrong…Maybe I'm just projecting or anthropomorphizing" [9]. I think, as do the majority, that this was indeed the case and that Lamoine's declaration was not particularly scientific. *Nevertheless*, this is what Lamoine was led to believe. And it isn't implausible to suggest that others will too – on the Replika website there is even a page dedicated to reassuring users their *Replika* isn't sentient, even though it might seem like it is [10].

What does anthropomorphisation have to do with AI *Hype*? In our call for papers, we roughly define AI Hype as the misrepresentation and miscommunication of AI's present capabilities and performance. We suggest that this is concerning because the overinflation of what AI *is* undermines meaningful public discourse around it. What I want to show here is that the anthropomorphisation of AI contributes to this widespread misrepresentation and miscommunication by inducing ideas of equivalence, superiority, and inferiority *relative to humans*. In other words, the anthropomorphisation of automated systems implicates humans as a benchmark for comparison, but these comparisons aren't fit for purpose. The object of this article is to explicitly point to examples of this phenomenon, explain how they contribute to hype, and consider how to mitigate them (if we should at all).

Terminology

The words used to describe particular topics, concepts, and phenomena affect how we think about them. Even initially implausible rhetoric can begin to sound true and fluent once we process it multiple times [11]. It is particularly important, therefore, that the words we use to communicate about important topics – such as AI – are accurate.

Anthropomorphisation impedes this objective. Whilst natural, using human-like terms to communicate about AI leads to misleading metaphors, similarities, and tropes.

One recent example is Open AI's latest 'upgrade' to ChatGPT: that it "can now see, hear, and speak" [12]. What they really mean is that ChatGPT can now recognise and analyse pictures, transcribe speech, and respond with a voice instead of just text. But this is far less catchy. The problem though is that attributing sensory modalities to an artificial system is, at best, clearly false to most people, and at worst, implies a capacity for experience. In the ordinary use of the term, when we say we saw something we are referring to the experience of perceiving something. For instance, we might describe what *it was like* to perceive the redness of an apple. The same is true for audition. Speech is slightly different but implies a degree of cognitive capacity greater than *just* the capacity to produce text.

A similar complaint has been made for the use of 'hallucination'. Ordinarily, the term hallucination refers to when humans perceive phenomena that are not *really* there. But it has been reappropriated by AI discourse to refer to when LLMs, like ChatGPT, simply make up (or hallucinate) facts[1]. But again, this implies they have some capacity to perceive. Which they don't. As Emily Bender has rightly pointed out, 'hallucinate' "is a terrible word choice…suggesting as it does that the language model has *experiences* and *perceives things*" [14].

These are examples of using *verbs* to anthropomorphise AI. And there are plenty more. It is common to see newspaper headlines claim AI is 'understanding', 'thinking', 'confusing', or 'going rogue' and 'misbehaving', for example. We also use anthropomorphised *adjectives* when talking about AI. Chatbots like Replika, for instance, sometimes portray particular personalities like 'friendly', 'kind', 'devious', and 'attractive'. They are also spoken about as if they have particular emotions like being 'sad', 'angry', 'happy', or 'horny'. But do chatbots really have personalities in the ways humans do? Do they have a capacity for emotion? No. Just like it seems misleading to ascribe our pet rock with happiness, so it seems misleading to ascribe a chatbot with happiness. 'Intelligence' is another hype-entangled adjective - some even argue the term AI itself is

[1] The term 'hallucination' was used at least 3 years before even the advent of GPT-2 [13] and only used in this way within computer science communities. Unfortunately, the term is now widely used without clarification.

hype [15]. Indeed, a buzzterm among certain AI circles are claims or goals of *human-like* intelligence and eventual *super*intelligence. Human-like intelligence is a clear comparison by equivalence. And *super*intelligence, broadly defined as a system infinitely smarter than humans, is a clear example of comparison by superiority.

Anthropomorphised terminology can be both symptomatic and inducive of hype. And when we refer to AI using human-like terms, we project misplaced beliefs onto such systems. That they *possess* human-like capabilities, emotions, and reasoning. This leads to overestimations of an AI's ability. Because public AI literacy is so low[2], by using such anthropomorphic language people might be more likely to believe AI systems really are capable of the tasks and processes such terms suggest. This then plays into overblown fears about job losses and, within policy circles, overconfidence about deploying algorithms for significant tasks like police facial recognition.

Imagery

Perhaps more important than the words used to communicate about AI, are the images. If you search 'AI' in Google Images, it is likely that among the many tropes presented (glowing brains, variations of *The Creation of Adam* [18], the colour blue, matrix style descending code, and terminators) you will likely come across a (white [19]) human-looking robot. As Dihal and Duarte explain [20]:

> Pictures that show humanoid robots in deep contemplation, or tackling difficult maths problems on a blackboard, reinforce unrealistic fears and expectations about AI achieving human-like intelligence, or even 'superintelligence', imminently. This overshadows current concerns about AI and overhypes present capabilities. AI does not 'think', it is a programme executing algorithms.

The use of such images, again, feeds human-like comparisons and plays into general public fears and misunderstandings about AI. These misunderstandings and misrepresentations can be especially propagated through anthropomorphised imagery because of their association with popular culture and science fiction. Terminator images,

---

[2] The CDEI's December 2021 public attitudes to AI tracker, for example, found that only 13% felt they had a strong understanding of AI [16]. For a more recent study, see [17].

commonly used in articles about AI, spread narratives of fear and uprising independently of the article's content. These narratives feed into overhyped claims of superintelligence, making them seem more plausible through association [20].

Interestingly, this narrative of comparison has been used by some to reiterate human exceptionality. Since 2022, the UK Army has been running a TV advert titled: 'The Army of the Future'. It begins by showing a Terminator-type robot in a battleground, fading to a human soldier with the accompanying narration: "What does the army of the future look like? It looks like you." [21].

Morality

Recall Lamoine. Lamoine's interactions with LaMDA led them to believe that LaMDA was sentient. You might think this ascription belongs to the section on 'terminology'. But the attribution of *sentience* to any entity, let alone an artificial one, is not like saying it comes across happy, or sad, or horny. It is a distinctly *moral* property. Sentience, often understood as the capacity for valenced phenomenal experience (or, in other words, to have bad and good experiences), seems to many to matter morally. This is what Lamoine thought. Not only did they declare that LaMDA was sentient, but they, because of this ascription, urged their colleagues to "treat LaMDA well" [8].

Philosophers sometimes refer to an entity with this special moral status as a moral *patient* [22, 23]. A moral patient is an entity that is susceptible to certain moral harms and benefits and is, as a result, owed particular duties and obligations. Moral *agents* are those bound by these duties and obligations. You and I are moral *agents*. Whether artificial systems are too is up for debate. You and I are also moral *patients*. If LaMDA really is sentient, then (for many, but by no means everyone) LaMDA is too. Similar reasoning guides our treatment of animals. For instance, the UK's Animal Welfare Act [24] turns crucially on notions of suffering [25].

But for many, this is absurd [26]. How could a human-made artefact ever deserve moral consideration? Joanna Bryson has claimed that ascribing robots moral patiency would "break everything – society, all ethics, all our values" [27]. Indeed, if it is the case that artificial entities *can* be moral patients, there will be an upheaval in society and law akin to animals.

It is paramount, however, that regardless of how this debate concludes, it isn't determined by anthropomorphic projection. It is likely that many, just like Lamoine did,

will anthropomorphise their AI (or robot) companions and interpret their human-like characteristics as indicators of sentience. But without *scientific* indicators, these ascriptions will be both symptoms and propagators of hype. Indeed, concepts like 'self-awareness' are heavily connected to sci-fi narratives of robot uprisings and human extinction.[3]

Linking AI Hype and Anthropomorphism

What is the explicit link between anthropomorphism and AI Hype? My answer is this: it limits our understanding of AI to *human terms*. Because of this, the mental models we create resemble, and base themselves on, human qualities. And, as a result, we end up working with a scale of non-human, human, and post-human. But these are inadequate simplifications. They lead to overblown hype around AI because human-like terms are inadequate and misleading.

But as I started off by saying: anthropomorphism is unavoidable. Indeed, some go so far as to say that instead of being a bug, our tendency to anthropomorphise is a *feature* of human nature [28, 29]. One of the many reasons anthropomorphisation is useful is that it functions to transcribe non-human behaviour and capacities into behaviours and capacities we relate to and understand. Because AI is something we do not understand it is perhaps natural for us to try and comprehend it in terms we *do* understand: human. Talking about and projecting human capacities onto non-human entities allows us to relate to and translate their distinctly inhuman behaviour.

The question is: ought we to do this? On the one hand, the anthropomorphisation of AI leads to misrepresentational hype because human-like terms do not adequately capture certain complexities. This simplification can lead to overblown fears around AI, not least because we end up thinking in terms of replacement. But on the other, we might rightfully ask what the alternative is. Anthropomorphisation allows us to comprehend and relate to a technology we fundamentally do not understand. Whilst this may lead to disanalogies, at least there is an element of interpretation.

---

[3] If I am right that these anthropomorphic projections ought not to determine moral consideration, then theories like John Danaher's [22] and Henry Shevlin's [23] that argue we only need to *perceive* certain morally relevant properties in artificial entities, rather than determine that they *actually* have them, run into problems.

I have mentioned the lack of public AI literacy. Anthropomorphic language and imagery, accompanied by many other factors like sci-fi narratives, clearly impedes true public understanding. But if anthropomorphism is unavoidable, perhaps before the damaging effects of anthropomorphic representations can be mitigated, AI literacy needs to be improved. Indeed, one might hope that as more of the public learns to spot inaccuracies in anthropomorphic terminology and imagery, wider media outlets will seek to avoid them for fear of misrepresentation charges. (One motivation they have now is that anthropomorphic terms are eye-catching; they lead to clicks and engagement). Perhaps, just as we realise ascribing happiness to our pet rock is wrong but not irresponsible, when public literacy around AI increases, using anthropomorphic terms to simplify and describe AI will be wrong but not irresponsible. This would be because such misrepresentation is *recognised* as being wrong and wouldn't, therefore, spread misinformation because the public would not be susceptible to it. Perhaps.

Concluding Thoughts

Here, I've illustrated some of the ways the anthropomorphism of AI feeds into AI Hype. In particular, I have surveyed its connection through language, imagery, and morality. I then concluded by considering what we ought to do about this phenomenon. In light of the seeming inevitability that humans will anthropomorphise AI (and other beings) even if they ought not to, I have painted a (perhaps idealistic) picture between public understanding and the disenfranchisement of inaccurate rhetoric. Rather than try to stop it, I have suggested that the anthropomorphisation of AI needs to be accommodated and its consequences mitigated.

Whether a truer public understanding of AI would achieve this, and whether AI literacy could even play catch up to such an extent, is questionable. A short-term solution is to acknowledge, alongside their use, that anthropomorphic characterisations of AI are inaccurate and to explain why. Indeed, in doing so, we would also be increasing public AI literacy and contributing to this long-term idealistic proposal.

References

1. Epley, N., Waytz, A., Cacioppo, JT.: On seeing human: A three-factor theory of anthropomorphism. Psychological Review (2007). https://doi.org/10.1037/0033-295X.114.4.864

2. Deroy, O. The Ethics of Terminology: Can We Use Human Terms to Describe AI? Topoi (2023). https://doi.org/10.1007/s11245-023-09934-1

3. Salles, A., Evers, K., Farisco, M. Anthropomorphism in AI. AJOB Neuroscience (2020). https://doi.org/10.1080/21507740.2020.1740350

4. Shneiderman, B., Muller, M. On AI Anthropomorphism. Medium. https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae (2023). Accessed 23 October 2023

5. Coeckelbergh, M. Three Responses to Anthropomorphism in Social Robotics: Towards a Critical, Relational, and Hermeneutic Approach. Int J of Soc Robotics (2022). https://doi.org/10.1007/s12369-021-00770-0

6. van Wynsberghe, A. Social robots and the risks to reciprocity. AI & Soc (2022). https://doi.org/10.1007/s00146-021-01207-y

7. Tong, A. What happens when your AI chatbot stops loving you back? Reuters. https://www.reuters.com/technology/what-happens-when-your-ai-chatbot-stops-loving-you-back-2023-03-18/ (2023). Accessed 23 October 2023

8. Tiku, N. Google engineer Blake Lemoine thinks its LaMDA AI has come to life. The Washington Post. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/ (2022). Accessed 23 October 2023.

9. Lamoine, B. Is LaMDA Sentient? — an Interview. Medium. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917 (2022). Accessed 23 October 2023

10. Is Replika sentient? Replika. https://web.archive.org/web/20230531013738/https://help.replika.com/hc/en-us/articles/360058852132-Is-Replika-sentient- (2023). Accessed 23 October 2023

11. Dechêne, A., Stahl, C., Hanse, J., Wänke, M. (2010) The Truth About the Truth: A Meta-Analytic Review of the Truth Effect. Personality and Social Psychology Review (2010). https://journals.sagepub.com/doi/10.1177/1088868309352251. Accessed 23 October 2023

12. ChatGPT can now see, hear, and speak. OpenAI. https://web.archive.org/web/20231021033809/https://openai.com/blog/chatgpt-can-now-see-hear-and-speak (2023). Accessed 23 October 2023

13. Hariharan, B., Girshick, R. Low-shot Visual Recognition by Shrinking and Hallucinating Features. Proceedings of the IEEE international conference on computer vision. https://openaccess.thecvf.com/content_ICCV_2017/papers/Hariharan_Low-Shot_Visual_Recognition_ICCV_2017_paper.pdf (2017).

14. Bender, Emily. [Tweet]. https://t.co/oIgCZYOnSM (Nov 1, 2022). Accessed 23 October 2023.

15. Raji, ID. AI's Present Matters More Than Its Imagined Future. The Atlantic. https://www.theatlantic.com/technology/archive/2023/10/ai-chuck-schumer-forum-legislation/675540/ (2023). Accessed 23 October 2023

16. Public Attitudes to Data and AI - Tracker Survey. CDEI. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1092140/Public_Attitudes_to_Data_and_AI_-_Tracker_Survey.pdf (2022). Accessed 23 October 2023.

17. Brauner, P., Hick, A., Philipsen, R., Ziefle, M. What does the public think about artificial intelligence? - A criticality map to understand bias in the public perception of AI. Frontiers in Computer Science. https://www.frontiersin.org/articles/10.3389/fcomp.2023.1113903/full (2023).

18. Singler, B. The AI creation meme: A case study of the new visibility of religion in artificial intelligence discourse. Religions. https://doi.org/10.3390/rel11050253 (2020).

19. Cave, S., Dihal, K. The whiteness of AI. Philosophy & Technology. https://doi-org.10.1007/s13347-020-00415-6 (2022).

20. Dihal, K., Duarte, T. Better Images of AI: A Guide for Users and Creators. Better Images of AI. https://blog.betterimagesofai.org/wp-content/uploads/2023/02/Better-Images-of-AI-Guide-Feb-23.pdf (2023). Accessed 23 October 2023.

21. Army Jobs. The Army of The Future. YouTube. https://www.youtube.com/watch?v=fIeO03BdNq4 (2022). Accessed 23 October 2023.

22. Danaher, J. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. Sci Eng Ethics. https://doi.org/10.1007/s11948-019-00119-x (2022).

23. Shevlin, H. How Could We Know When a Robot was a Moral Patient? Cambridge Quarterly of Healthcare Ethics. https://doi.org/10.1017/s0963180120001012 (2022).

24. Animal Welfare Act 2006. https://www.legislation.gov.uk/ukpga/2006/45 (2006). Accessed 23 Oct 2023

25. Bentham, J. The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation. Clarendon Press (1996).

26. Birhane, A., van Dijk, J. Robot rights? Let's talk about human welfare instead. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3375627.3375855 (2020).

27. Bryson, J. One Day, AI Will Seem as Human as Anyone. What Then? Wired. https://www.wired.com/story/lamda-sentience-psychology-ethics-policy/ (2022). Accessed 23 Oct 2023

28. Gunkel, D. Duty Now and for the Future: Communication, Ethics and Artificial Intelligence. Journal of Media Ethics. https://doi.org/10.1080/23736992.2023.2264854 (2023).

29. Darling, K. The New Breed: How to Think About Robots. Penguin, London (2021).