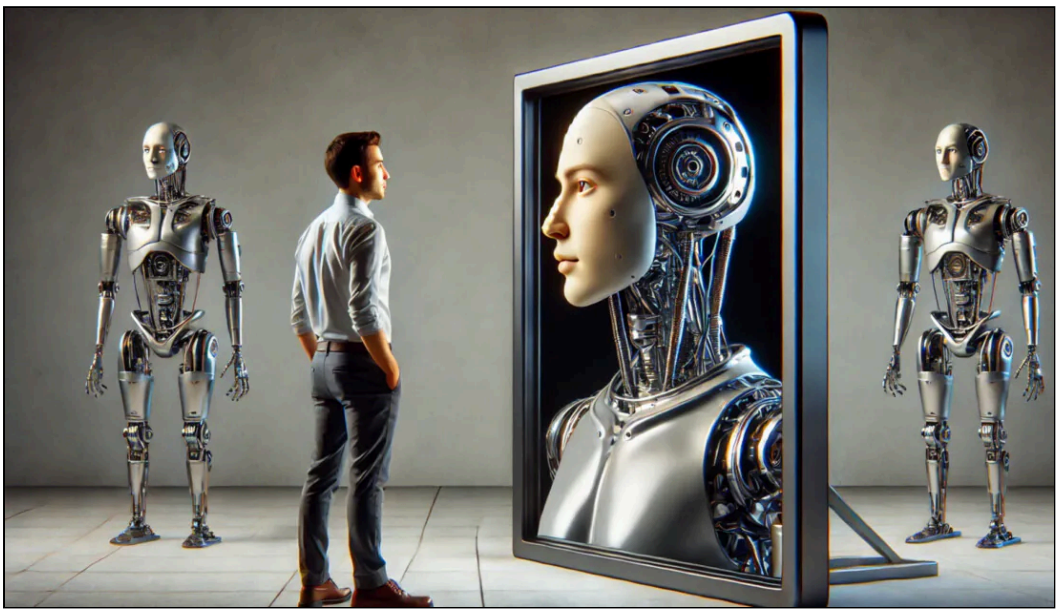
A Call to Moderate Anthropomorphism in AI Platforms

Originally published at https://www.unite.ai/a-call-to-moderate-anthropomorphism-in-ai-platforms/



OPINION Nobody in the fictional *Star Wars* universe takes AI seriously. In the historic human timeline of George Lucas's 47 year-old science-fantasy franchise, threats from singularities and machine learning consciousness are absent, and AI is confined to autonomous mobile robots (*'droids'*) – which are habitually dismissed by protagonists as mere 'machines'.

Yet most of the *Star Wars* robots are highly anthropomorphic, clearly designed to engage with people, participate in 'organic' culture, and use their simulacra of emotional state to bond with people. These capabilities are apparently designed to help them gain some advantage for themselves, or even to ensure their own survival.

The 'real' people of *Star Wars* seem immured to these tactics. In a cynical cultural model apparently inspired by the various eras of slavery across the Roman empire and the early United States, Luke Skywalker doesn't hesitate to buy and restrain robots in the context of slaves; the child Anakin Skywalker abandons his half-finished C3PO project like an unloved toy; and, near-dead from damage sustained during the attack on the Death Star, the 'brave' R2D2 gets about the same concern from Luke as a wounded pet.

This is a very 1970s take on artificial intelligence*; but since nostalgia and canon dictate that the original 1977-83 trilogy remains a template for the later sequels, prequels, and TV shows, this human insensibility to AI has been a resilient through-line for the franchise, even in the face of a growing slate of TV shows and movies (such as *Her* and *Ex Machina*) that depict our descent into an anthropomorphic relationship with AI.

Keep It Real

Do the organic *Star Wars* characters actually have the right attitude? It's not a popular thought at the moment, in a business climate hard-set on maximum engagement with investors, usually through viral demonstrations of visual or textual simulation of the real world, or of human-like interactive systems such as Large Language Models (LLMs).

Nonetheless, a new and brief paper from Stanford, Carnegie Mellon and Microsoft Research, takes aim at indifference around anthropomorphism in AI.

The authors characterize the perceived 'cross-pollination' between human and artificial communications as a potential harm to be urgently mitigated, for a number of reasons †:

> *'[We] believe we need to do more to develop the know-how and tools to better tackle anthropomorphic behavior, including measuring and mitigating such system behaviors when they are considered undesirable.*
>
> *'Doing so is critical because—among many other concerns—having AI systems generating content claiming to have e.g., feelings, understanding, free will, or an underlying sense of self may erode people's sense of agency, with the result that people might end up attributing moral responsibility to systems, overestimating system capabilities, or overrelying on these systems even when incorrect.'*

The contributors clarify that they are discussing systems that are *perceived* to be human-like, and centers around the potential *intent* of developers to foster anthropomorphism in machine systems.

The concern at the heart of the short paper is that people may develop emotional dependence on AI-based systems – as outlined in a 2022 study on the gen AI chatbot platform Replika) – which actively offers an idiom-rich facsimile of human communications.

Systems such as Replika are the target of the authors' circumspection, and they note that a further 2022 paper on Replika asserted:

> *'[U]nder conditions of distress and lack of human companionship, individuals can develop an attachment to social chatbots if they perceive the chatbots' responses to offer emotional support, encouragement, and psychological security.*
>
> *'These findings suggest that social chatbots can be used for mental health and therapeutic purposes but have the potential to cause addiction and harm real-life intimate relationships.'*

De-Anthropomorphized Language?

The new work argues that generative AI's potential to be anthropomorphized can't be established without studying the social impacts of such systems to date, and that this is a neglected pursuit in the literature.

Part of the problem is that anthropomorphism is difficult to define, since it centers most importantly on language, a human function. The challenge lies, therefore, in defining what 'non-human' language exactly sounds or looks like.

Ironically, though the paper does not touch on it, public distrust of AI is increasingly causing people to reject AI-generated text content that may appear plausibly human, and even to reject human content that is deliberately mislabeled as AI.

Therefore 'de-humanized' content arguably no longer falls into the 'Does not compute' meme, wherein language is clumsily constructed and clearly generated by a machine.

Rather, the definition is constantly evolving in the AI-detection scene, where (currently, at least) excessively clear language or the use of certain words (such as '*Delve*') can cause an association with AI-generated text.

> *'[L]anguage, as with other targets of GenAI systems, is itself innately human, has long been produced by and for humans, and is often also about humans. This can make it hard to specify appropriate alternative (less human-like) behaviors, and risks, for instance, reifying harmful notions of what—and whose—language is considered more or less human.'*

However, the authors argue that a clear line of demarcation should be brought about for systems that blatantly misrepresent themselves, by claiming aptitudes or experience that are only possible for humans.

They cite cases such as LLMs claiming to 'love pizza'; claiming human experience on platforms such as Facebook; and declaring love to an end-user.

Warning Signs

The paper raises doubt against the use of blanket disclosures about whether or not a communication is facilitated by machine learning. The authors argue that systematizing such warnings does not adequately contextualize the anthropomorphizing effect of AI platforms, if the output itself continues to display human traits†:

> *'For instance, a commonly recommended intervention is including in the AI system's output a disclosure that the output is generated by an AI [system]. How to operationalize such interventions in practice and whether they can be effective alone might not always be clear.*
>
> *'For instance, while the example "[f]or an AI like me, happiness is not the same as for a human like [you]" includes a disclosure, it may still suggest a sense of identity and ability to self-assess (common human traits).'*

In regard to evaluating human responses about system behaviors, the authors also contend that Reinforcement learning from human feedback (RLHF) fails to take into account the difference between an appropriate response for a human and for an AI†.

> *'[A] statement that seems friendly or genuine from a human speaker can be undesirable if it arises from an AI system since the latter lacks meaningful commitment or intent behind the statement, thus rendering the statement hollow and deceptive.'*

Further concerns are illustrated, such as the way that anthropomorphism can influence people to believe that an AI system has obtained 'sentience', or other human characteristics.

Perhaps the most ambitious, closing section of the new work is the authors' adjuration that the research and development community aim to develop 'appropriate' and 'precise' terminology, to establish the parameters that would define an anthropomorphic AI system, and distinguish it from real-world human discourse.

As with so many trending areas of AI development, this kind of categorization crosses over into the literature streams of psychology, linguistics and anthropology. It is difficult to know what current authority could actually formulate definitions of this type, and the new paper's researchers do not shed any light on this matter.

If there is commercial and academic inertia around this topic, it could be partly attributable to the fact that this is far from a new topic of discussion in artificial intelligence research: as the paper notes, in 1985 the late Dutch computer scientist Edsger Wybe Dijkstra described anthropomorphism as a 'pernicious' trend in system development.

> *'[A]nthropomorphic thinking is no good in the sense that it does not help. But is it also bad? Yes, it is, because even if we can point to some analogy between Man and Thing, the analogy is always negligible in comparison to the differences, and as soon as we allow ourselves to be seduced by the analogy to describe the Thing in anthropomorphic terminology, we immediately lose our control over which human connotations we drag into the picture.*
>
> *'…But the blur [between man and machine] has a much wider impact than you might suspect. [It] is not only that the question "Can machines think?" is regularly raised; we can —and should— deal with that by pointing out that it is just as relevant as the equally burning question "Can submarines swim?"'*

However, though the debate is old, it has only recently become very relevant. It could be argued that Dijkstra's contribution is equivalent to Victorian speculation on space travel, as purely theoretical and awaiting historical developments.

Therefore this well-established body of debate may give the topic a sense of weariness, despite its potential for significant social relevance in the next 2-5 years.

Conclusion

If we were to think of AI systems in the same dismissive way as organic *Star Wars* characters treat their own robots (i.e., as ambulatory search engines, or mere conveyers of mechanistic functionality), we would arguably be less at risk of habituating these socially undesirable characteristics over to our human interactions – because we would be viewing the systems in an entirely non-human context.

In practice, the entanglement of human language with human behavior makes this difficult, if not impossible, once a query expands from the minimalism of a Google search term to the rich context of a conversation.

Additionally, the commercial sector (as well as the advertising sector) is strongly motivated to create addictive or essential communications platforms, for customer retention and growth.

In any case, if AI systems genuinely respond better to polite queries than to stripped down interrogations, the context may be forced on us also for that reason.

* *Even by 1983, the year that the final entry in the original* Star Wars *was released, fears around the growth of machine learning had led to the apocalyptic* War Games*, and the imminent* Terminator *franchise.*

† *Where necessary, I have converted the authors' inline citations to hyperlinks, and have in some cases omitted some of the citations, for readability.*

First published Monday, October 14, 2024

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More