https://doi.org/10.1038/s42254-023-00584-1

How to edit anthropomorphic language about artificial intelligence

Advances in artificial intelligence (AI) and the hype they are generating have raised concerns about how scientists should talk about AI systems. Here is how we will approach the editing of such language to ensure clarity, accuracy and avoid misinterpretation and anthropomorphism.

The fast-paced developments in AI are already impacting science both for the better, by providing powerful tools for discovery, and for the worse, by challenging the rigour of the scientific method and the trust in it. In a Viewpoint in this issue, we asked AI ethics and policy experts about the risks of using large language models (LLMs) and, more broadly, generative AI, in science. Their message is clear: the scientific community must deploy these tools responsibly and carefully consider their implications for good scientific practice. Furthermore, the Viewpoint authors call for researchers to collaborate with journals, publishers, conference organizers, AI ethics and safety experts to develop best practices, standards and policies to ensure that the benefits of using AI in research are balanced against the risk of fundamentally undermining science and its role in society. What can we, as a Nature Portfolio journal specifically, do to help?

Anthropomorphic language is widespread in physics: masses 'feel' the gravitational potential, photons 'know' the state of their entangled partner and spins generally 'want' to align. This language is fine because it's hard to associate human-like feelings, beliefs, or intent with most concepts in physics. However, that is no longer the case when we talk about AI systems. In this context, the use of "rich psychological terms"[1] (which include awareness, perception, agency, understanding, knowledge and theory of mind) requires careful consideration because it can "impair scientific communication and understanding and invite premature conclusions of ethical or legal significance"[1]. With the impressive capabilities of LLM chatbots, it's hard not to see human-like characteristics, but we should all strive to "resist the siren call of anthropomorphism"[2]. That's where editors can help.

While handling AI-related articles in the past we have also succumbed to the siren call and for this reason we are developing an approach to keep us safe when we sail into murky waters. First, we will try to avoid at all costs the use of 'the AI/an AI' due to its unfortunate suggestion of agency. Instead, we will either change to 'the AI system/an AI system' or be very clear what we are talking about. Are we referring to a generic type of deep learning model, as in 'a generative model'; a particular type of generative model, as in 'a LLM' or 'a diffusion model'; or to a specific instance, as in a GPT-3 model or a LaMDA model; or perhaps chatbots built using these, such as ChatGPT or Bard? Being specific will hopefully help the reader gauge what can be expected from the subject of the sentence.

> "We should all strive to 'resist the siren call of anthropomorphism'."

Rich psychological terms should be avoided, but that will not always be possible, so they should at least be defined and justified. When assessing whether the use of a term is appropriate it's worth asking what the AI system actually does. For example, a LLM generates "statistically likely continuations of word sequences"[2], so a LLM chatbot cannot 'know', 'think', 'claim' or 'suggest' in the sense a human can. We should then consider whether there is any danger that the term is misunderstood or misinterpreted in the given context. In the above example 'claim' and 'suggest' imply intent, whereas 'know' and 'think' cognition. If that is the case, then is their use justified or can the term perhaps be replaced by a more precise word? If that is not possible, then we suggest considering how the meaning or usage of the term in the particular context can be clarified or defined. When nothing else works we'll use quotation marks to emphasize the abuse of the term.

The above guidance is not a journal policy, but a framework to help editors approach the issue of anthropomorphic language in a systematic way. Although it should help us deal with the most common misuses, it cannot be exhaustive, and we will likely be continuously refining this guidance as we encounter new examples. As in all our editing, we are striving for clarity, accuracy and avoiding possible misinterpretation, and we can only do this if our authors are on board. So, we call on our community for feedback to help us develop and disseminate good scientific communication practices.

We thank Bryan Kaiser for raising this issue and Murray Shanahan, Jenn Richler and Liesbeth Venema for useful discussions.

Published online: 26 April 2023

References

1. Shevlin, H. & Halina, M. Apply rich psychological terms in AI with care. *Nat. Mach. Intell.* **1**, 165–167 (2019).
2. Shanahan, M. Talking about large language models. Preprint at https://arxiv.org/abs/2212.03551 (2023).