What Do We Mean When We Say "Anthropomorphism"?

Dorian Liu[1] and Cathy Li[2]

Abstract. "That is just anthropomorphism" is among the most frequent rebuttals in current debates about machine minds, and it is standardly used to end the discussion. We ask what the charge asserts, and when it is legitimate. Using Clever Hans as a running example, we analyse an attribution of mind into four elements: a subject S, an attributed property P, the conditions C that P requires, and the indicator f on which the attributor relies. A legitimate charge of anthropomorphism, we argue, requires the accuser to produce a trivial competing explanation of the indicator. We then identify three ways its prominent uses fail to do so. Finally, we argue that only with the charge set aside can we see what these disputes are really about.

1 Introduction

"That is just anthropomorphism." In debates over whether machines have minds, this is among the most common replies, and it is normally meant to settle them; yet whether the charge holds, or is merely asserted, is rarely examined. In its general form, it says that "we project human-like qualities onto other things on the basis of only superficial similarities" [1].

In practice, however, a charge of anthropomorphism usually carries two claims at once. The first is an epistemic claim about the attributor: the attribution is a projection, arising from a disposition in us and resting on mere resemblance, owing nothing to the evidence. The second is an ontological claim about the object: that the object does not have the property. We therefore want to ask what we are really claiming when we make a charge of anthropomorphism, and whether the move from the epistemic claim to the ontological one is legitimate.

A reasonable objection is that the psychology of anthropomorphism is by now well documented, so that the concept needs no further clarification [2, 3]. But that psychology concerns only the first of the two claims, how an attribution is produced; it says nothing about whether the move from the first claim to the second is a good one. Nor is the conceptual question settled in philosophy: as we show in Section 3, even the most recent scholarly treatments fail to keep the two claims apart.

The paper is as follows. In Section 2 we map the charge onto four variables and ask what a legitimate charge of anthropomorphism requires. In Section 3 we test its recent academic and public uses against that standard, and examine what is perhaps the most widely circulated charge of all, that the system "is just predicting the next token". Section 4 concludes, drawing from the analysis a short set of requirements that both parties to the dispute must meet.

[1] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, email: zhaoting.liu6@gmail.com
[2] School of Mathematics, University of Edinburgh

2 What a Legitimate Charge Requires

The charge of anthropomorphism is far from a creature of the language-model age, and perhaps its most famous occasion is a horse. Around 1900 Clever Hans was exhibited across Germany as an animal that could do arithmetic. Posed a sum aloud, he would strike out the answer with his hoof. The demonstrations were public and repeatable, and no manifest trickery could be detected, so that almost everyone drew the same conclusion: Hans had a share of the mathematical capacity we credit to ourselves.

Suppose you had been in the audience and had drawn the natural conclusion. Someone beside you objects: you are anthropomorphising; the horse has no mathematical ability. In what sense does this objection hold? To see its structure, some heuristic formalisation might help. From the charge we can separate four variables:

- S, the subject, the thing to which the property is credited; here, the horse.
- P, the property attributed to S; here, arithmetical ability.
- C, the condition S must satisfy to have P; here, that the horse is really calculating.
- f, the evidence the attributor goes on; here, that Hans strikes the right number.

We can then set out two competing chains of inference. Take the attributor first. His inference can be set out as follows:

(i) f holds, that is, Hans strikes the right number.

(ii) f holds only where C holds, that is, only a horse that is really calculating strikes the right number.

(iii) So f establishes that S satisfies C, and thereby supports P.

(iv) Therefore Hans can do arithmetic.

Now take the objector:

(i') f holds, that is, Hans strikes the right number.

(ii') f holds in cases where C does not, that is, a horse that cannot calculate may still strike the right number.

(iii') So f does not establish that S satisfies C, and might not support P.

(iv') The attributor nonetheless holds that f supports P, and does so by projection.

The disagreement turns on (ii') and (iv'). The first concerns the horse and its evidence: whether f can hold where C does not. The second concerns the attributor: with f no longer supporting P, whether his holding to P is the work of projection. We take them in turn.

Take the first. To attribute arithmetic to Hans is to read his tapping as a sign that he is really calculating. To make the objection, the objector must point to some other way the right taps could come about; that is, the objector needs a possible competing explanation. Suppose the questioner knows the answer. As Hans approaches the right number, the questioner unconsciously tenses, and relaxes the moment it is reached; Hans reads this slight change in the questioner's posture. On this account Hans is responding to the questioner. If that is what is happening, the taps no longer show that Hans is calculating, and so give no support to the claim that he can do arithmetic.

A merely possible competing explanation, however, is not enough. Suppose the accuser proposes that Hans understands German: he recalls the answer he was taught, and taps it out. This does not require the horse to calculate, so it does not require S to satisfy C. Yet it leaves the attribution no less anthropomorphic, because what it requires of the horse, a capacity for language, is a human capacity as demanding as arithmetic itself. Crediting Hans with German is, in this sense, no more credible than crediting him with arithmetic; the accuser has merely relocated the projection. So we need to place a further constraint on the competing explanation: it must be trivial, in that the condition its operation requires of S must be weaker than C, so that an S in which it operates need not satisfy C and need not have P. In this sense attention to posture is trivial, while understanding German is not.

Two further requirements may also be necessary. First, the competing explanation must be defensible: the accuser cannot merely say that some other mechanism must be at work, or that such a mechanism is conceivable, but must have evidence, independent of f, that it is real and operative. Second, it must not already have been excluded by the attributor: if he has good reason to rule it out (he has screened the questioner off, say, and the horse still strikes the right number), it no longer unsettles f's support for P.

The first condition of a legitimate charge is therefore met when a competing explanation can be exhibited that produces f without requiring S to satisfy C, is trivial in the sense given, comes with justifying evidence, and has not been excluded by the attributor.

The second condition of a legitimate charge, that the attribution is the work of projection, is of a different kind: it concerns why the attributor holds that f tracks P. As a claim about the source of his belief, appeal to a psychological fact (whether an act of projection actually occurred, say) is not operational, since he may well deny it, just as the objector may vaguely gesture at some indefensible competing hypothesis. We do better to read it off his epistemic situation. Suppose there is a trivial competing explanation that produces f without the horse's working out the sum as we do. Then two typical cases arise. In the first, the attributor knows of the competing explanation and still holds that f tracks P. In the second, he does not, and supposes that only a horse working out the sum as we do could strike the right number. Yet whether or not he knows, he could have withheld the attribution and not credited the horse with P. That he attributes it all the same is best understood as projection: his confidence in P is sustained by the subject's likeness to us.

A legitimate charge of anthropomorphism thus has two conditions. The first falls on the accuser, who must be able to offer an explanation of f that is trivial, defensible, and not yet excluded by the attributor, showing that f does not in fact support P. The second falls on the attributor: with f thus failing to support P, his confidence in P is sustained by the subject's likeness to us, which is to say by projection. A charge of anthropomorphism is legitimate just when both conditions hold.

It bears emphasis that f need not be outward behaviour, and that this is what the charge turns on.[3]

3 The Charge in Use

Section 2 fixed the reach of a legitimate charge: at most it establishes the epistemic claim, that f does not support P; it does not establish the ontological claim, that S lacks P. In practice the two are run together, and that conflation is where the charge does its damage. We distinguish three ways its prominent uses go wrong.

The first failure is overreach: the charge establishes only the epistemic claim yet is used to assert the ontological one as well. Perhaps the clearest public instance is Ted Chiang's recent essay [4], whose argument can be reconstructed as follows.

(P1) The system produces fluent, humanlike text (f).

(P2) There is a trivial competing explanation of f: the text is to be taken as "a deepfake medium", on the grounds that generating a plausible simulacrum of a conversation between conscious beings is far easier than building a program that is genuinely conscious, and such a simulacrum does not invoke consciousness.

(C1) So f does not support attributing consciousness to the system (f does not support P).[4]

(C2) So the system is not conscious (S lacks P).[5]

[3] It is natural to suppose that f must be outward behaviour, so that any attribution resting on behaviour is anthropomorphic and any resting on inner structure is safe; but neither half holds. Behaviour need not invite the charge: let f be that Hans strikes the right number when the questioner is screened off and ignorant of the answer. This is still behaviour, but the cue-reading explanation no longer fits it, since that explanation requires a questioner the horse can read; an attribution on this f is not anthropomorphic, whatever else may be said against it. Inner evidence is not exempt either: someone who credits a network with understanding because its wiring diagram resembles a cortex goes on f that is internal, yet the resemblance is a competing explanation that has not been ruled out. What the charge turns on is whether an explanation that does not invoke P stands unexcluded; whether f is outward or inner does not by itself decide it.

[4] That is, since the trivial competing explanation has not been ruled out, f is equally compatible with the system's being conscious and with its merely simulating, and so does not discriminate between them.

[5] C2 is Chiang's stated conclusion [4].

From (P1) and (P2), only (C1) follows legitimately. To reach (C2) one would have to test the competing explanation itself and examine the system, not merely note that f falls short, and that is the step Chiang does not take.

The same conflation enters the scholarly definitions. Placani [5], for instance, characterises anthropomorphism as "a distinctively human process of inference or interpretation", and yet also classifies it 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". Anthropomorphism cannot be both at once. As an epistemic matter it says that f does not support P; a "factual error" is the ontological claim, asserting that the object in fact lacks the feature. If an attribution does saddle an object with a feature it lacks, that is a mistake the user has made; it does not fall under the concept of anthropomorphism. To list a use-level error (the factual one) alongside anthropomorphism (an epistemic one) as its two branches is a category mistake.

The second failure is that the charge does not apply to every attribution, because it presupposes that the accused has attributed P on the strength of f, the system's humanlike performance.

(P1) The accused attributes P to the system.

(P2) What he attributes P on is f, the system's humanlike performance.

(C) And f is taken to support P only by way of projection, reading "resembles us" as "has P"; so the attribution is unsupported.

The charge finds its target only where (P2) holds. But some of the accused go on something other than f. Take Hinton [6], whose argument can be reconstructed as follows.

(P1') When someone says "I have a subjective experience of X", the function of the words is to report that their perceptual system has gone wrong, by saying how the world would have to be for the perception to be correct. Talk of subjective experience is, on this account, a way of reporting the gap between how things are and how one perceives them.

(P2') A multimodal model with a camera, fooled by a hidden prism, points in the wrong direction; once told about the prism, it says that the object is really in front of it and that it had the subjective experience of the object being off to one side. The model is reporting exactly that gap.

(C) So the model uses the words "subjective experience" as we use them, and therefore has subjective experience.

This argument rests throughout on a semantic analysis of "subjective experience" and on the prism experiment; f, whether the outputs read as humanlike, never figures in it. So (P2) does not hold, and the charge of projection misses: it attacks a premise the accused never offered. The fitting response is to test the argument itself, challenging the redefinition of "subjective experience" or asking whether the model's report really means what a human's does, whereas dismissing it as anthropomorphism simply sidesteps the reasons actually given.[6]

The third failure occurs at the first threshold, whether the competing explanation succeeds at all. Recall the requirement of Section 2: for a competing explanation to deprive f of its support for P, what it requires of the system must be weaker than C, since only then can a system that merely satisfies it fail to satisfy C, and so need not have P. Cue-reading works precisely because what it requires (reading posture) is far weaker than C (calculating).

In the AI debate the competing explanation most often offered is "it is just predicting the next token". All of its force rests on an unstated premise: that predicting the next token is something undemanding, far weaker than C (whether C is understanding, a world model, or something else), as reading posture is weaker than calculating. If that were so, then f (the system answering fluently, solving new problems) would be fully compatible with the system's merely predicting, and f would be no reliable indicator of P. The claim looks intuitive, but it stands entirely on that premise, that predicting the next token requires very little.

That premise, at least to us, is doubtful. The cleanest way to test what predicting this well requires is to take a system least likely to harbour rich internal structure: an early, small, thoroughly un-humanlike model, such as Othello-GPT [7]; it is trained only to predict the next legal move in games of Othello, never told the rules and never shown a board. If predicting the next token (here the next move) really required very little, nothing close to C should be needed inside it. Yet under intervention, editing its internal representation so as to flip one square of the represented board, its subsequent judgments about which moves are legal change to match the edited board.[7] This suggests that to predict moves at this level the system must build, and causally use, an internal model of the board. Predicting the next move, then, cannot confidently be said to require anything weaker than C.

So this competing explanation is not the plainly trivial thing it is taken to be, and it therefore cannot serve the accuser as a competing explanation at all; in that sense the charge fails. As noted above, none of this shows that the system cannot understand; it shows only that "it is just predicting the next token" can no longer be taken as a free competing explanation, one that cancels f at no cost. For the charge to be legitimate, someone would first have to argue that predicting the next token at the frontier level does require less than C, which, on the evidence above, is hard to sustain.

[6] It is telling that Hinton is rarely accused of anthropomorphism, even as he ascribes subjective experience to machines. Part of the reason is surely that, as someone who understands the systems from the inside, he is plainly not reasoning from surface resemblance, so the charge, which targets inference from f, finds no purchase. One can imagine someone else advancing the very same argument and being dismissed as an anthropomorphiser all the same. The charge often turns on who is making the attribution, which shows again how narrow its legitimate range really is.

[7] Decodability alone would not settle the matter, since a probe can recover information a model carries but does not use [8, 9]; what settles it is the intervention, which shows the represented board is causally in use [10].

4 Conclusion

This paper takes no position on whether machines have minds, in part because that question is already well served by a large literature, and in part because, among philosophers themselves, such attributions rarely draw the charge of anthropomorphism at all. What we have offered is a way of taking apart the question that "that is just anthropomorphism" closes too early. The charge does real damage because it is present in almost every public debate while doing far less argumentative work than it appears to. Once it is set aside, the question can be brought down to the variables that actually carry it: which subject S is in question, which property P, what conditions C possessing P requires, and whether the evidence f excludes the competing explanations.

Two heuristic requirements follow. The first we may call conceptual transparency. Anyone who deploys the charge must say which subject S is in question,[8] which property P is at issue,[9] and which conditions C that property is taken to require: whether suffering, say, demands phenomenal consciousness, or whether understanding demands more than functional organisation. A charge that invokes what a system "really" understands or "genuinely" feels, while declining to state what reality here requires, merely presupposes its conclusion.

The second is symmetry of burden. Accuser and attributor alike must produce evidence or argument that their own position is not the naive one: at a minimum, the accuser must rule out the trivial competing explanation, and the attributor must offer substantive theoretical support; surface resemblance alone will not do. "S lacks P" is itself a claim about S, and it answers to C and f exactly as "S has P" does. If interpretability evidence reveals internal structure of the kind a mental capacity would require, that evidence cannot be waved away by reasserting that the system is "merely a machine", any more than behavioural resemblance alone settles the matter for the attributor.

Acknowledgements

We thank the AISB AICE reviewers for their comments on the extended abstract.

[8] The choice of S is easily overlooked yet often where a dispute really lies. The classic case is Searle's Chinese Room and the Systems Reply it provoked: Searle locates the subject as the person in the room, who understands no Chinese, while the reply relocates it to the whole system [11]. Clarifying S remains necessary for present systems; see Shanahan [12] on the distinction between the bare model and the system in which it is embedded, and Chalmers [13] on what kind of entity an LLM-based system is.

[9] A characteristic dispute over P can be seen in the recent public exchange on machine consciousness. Pope Leo XIV, in Magnifica Humanitas (2026), denies that AI systems are conscious, where to be conscious is to undergo experience, to have a body, to feel joy and pain, to mature through relationships. In the same period, interpretability work reports functional emotion in large models, emotion representations with causal roles in behaviour, while stating that "none of this tells us whether language models actually feel anything" [14]. The two do not contradict each other: they concern different properties P, with different conditions C, so an attribution of the one is consistent with a denial of the other. They appear to disagree only because they share a word.

REFERENCES

[1] Anil K. Seth. The mythology of conscious AI. Noema Magazine, 2026.
[2] Nicholas Epley, Adam Waytz, and John T. Cacioppo, 'On seeing human: A three-factor theory of anthropomorphism', Psychological Review, 114(4), 864–886, (2007).
[3] Stewart Elliott Guthrie, Faces in the Clouds: A New Theory of Religion, Oxford University Press, New York, 1993.
[4] Ted Chiang. No, artificial intelligence is not conscious. The Atlantic, June 3, 2026.
[5] Adriana Placani, 'Anthropomorphism in AI: hype and fallacy', AI and Ethics, 4(3), 691–698, (2024).
[6] Geoffrey Hinton. Will digital intelligence replace biological intelligence? Richard Gregory Memorial Lecture, University of Bristol, 2 June 2025, 2025. https://www.youtube.com/watch?v=utlQ8UBVPBI.
[7] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg, 'Emergent world representations: Exploring a sequence model trained on a synthetic task', in ICLR 2023; arXiv:2210.13382, (2023).
[8] John Hewitt and Percy Liang, 'Designing and interpreting probes with control tasks', in Proceedings of EMNLP-IJCNLP 2019, pp. 2733–2743, (2019).
[9] Yonatan Belinkov, 'Probing classifiers: Promises, shortcomings, and advances', Computational Linguistics, 48(1), 207–219, (2022).
[10] Neel Nanda, Andrew Lee, and Martin Wattenberg, 'Emergent linear representations in world models of self-supervised sequence models', in BlackboxNLP 2023; arXiv:2309.00941, (2023).
[11] John R. Searle, 'Minds, brains, and programs', Behavioral and Brain Sciences, 3(3), 417–457, (1980).
[12] Murray Shanahan, 'Talking about large language models', Communications of the ACM, 67(2), 68–79, (2024).
[13] David J. Chalmers, 'What we talk to when we talk to language models'. PhilArchive preprint, 2025.
[14] Nicholas Sofroniew, Isaac Kauvar, Jack Lindsey, et al. Emotion concepts and their function in a large language model. arXiv:2604.07729, Anthropic, 2026.