Yve Smith is human and can make mistakes, including about AI.

Yve Smith
School of Computer Science
CIS Building, Northumbria University
Ellison Place, Newcastle NE1 8ST
United Kingdom
yve.a.smith@northumbria.ac.uk

This paper is a provocation and, as such, is vulnerable to rigorous academic scrutiny. It presents a series of speculative LLM disclaimers as a potential avenue for experientially altering users' understandings of LLM systems, but these are not intended for real-world use. Disclaimers are a narrowly bounded example of such experiential interventions and should not be considered a complete solution. Claims made in this paper about the irrelevance of anthropomorphism, the implications of language as a mode of interface, and the flattening of embodied experience into text remain open to dispute, and may be unfounded, incomplete or misleading. Reading this paper may cause confusion, irritation, amusement, uncertainty or fatigue. Please exercise caution and consult reliable corroborating sources before taking any action following the reading of this paper.

large language models, anthropomorphism, metaphor, human-AI interaction, XAI

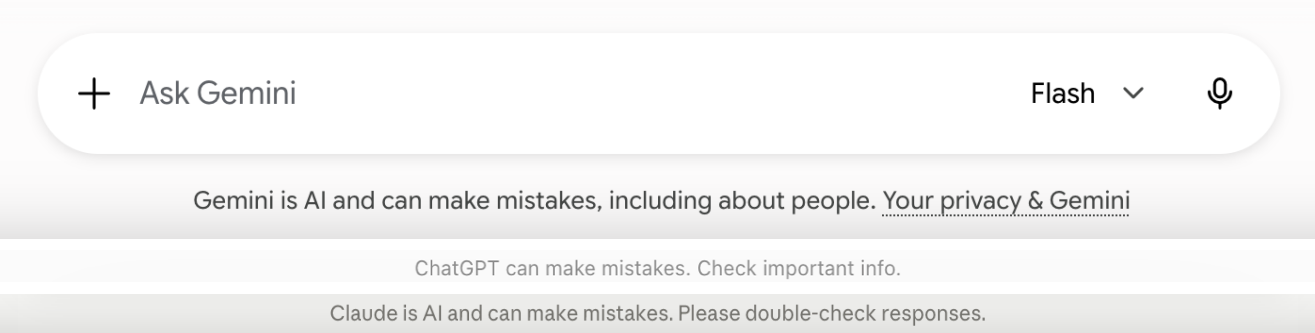


Figure 1: *Chat interface disclaimers for three of the major large language model (LLM) chatbots. They read "ChatGPT can make mistakes. Check important info.", "Claude is AI and can make mistakes. Please double-check responses." and "Gemini is AI and can make mistakes, including about people. Your privacy & Gemini."*

1. INTRODUCTION

Chatbots powered by large language models (LLMs) now serve as daily interlocutors for myriad users and functions, from professionals optimising workflows (Bick et al. 2026) to vulnerable individuals looking for emotional support (McBain et al. 2025). Their superficial similarity to a human on the other side of a chat window has led to widespread concern about the implications of our human tendency to anthropomorphise systems, combined with design decisions that amplify anthropomorphism in the name of user experience (Abercrombie et al. 2023).

However, a more pressing concern for the lay user is the opaqueness of their generative nature, and the need to decouple their linguistic fluency and apparent confidence with their actual reliability and accuracy (Bender and Koller 2020; Smith et al. 2023). As a measure to tackle this, AI disclaimers have become ubiquitous across many LLM chatbots (Cheng 2025); ever-present reminders of the system's fallibility (Figure 1).

© Smith. Published by BCS Learning and Development Ltd.

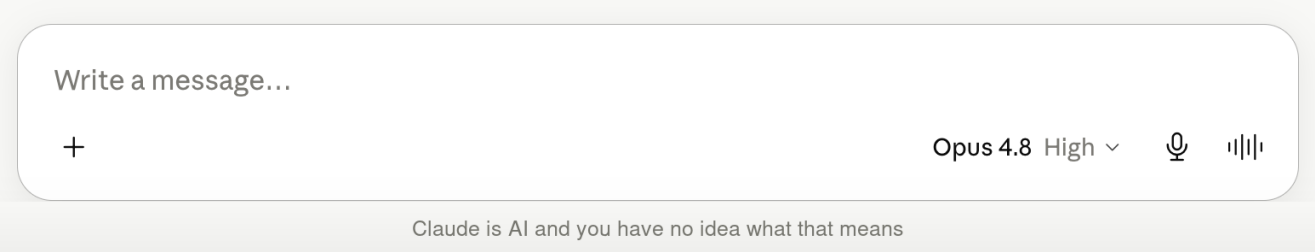


Figure 2: *Speculative chat interface disclaimer reading "Claude is AI and you have no idea what that means."*

In contrast to this conventional approach, this paper presents speculative disclaimers intended to confront and contend with the uncertainties that often go unaddressed in both existing disclaimers and the wider rhetoric about AI, using the disclaimer as an experiential vehicle through which to explore and interrogate how we might best encounter, introduce and explain AI systems.

2. ANTHROPOMORPHISM IS IRRELEVANT.

Whether it is being leveraged for usability or demonised as manipulation (Abercrombie et al. 2023; So et al. 2026), anthropomorphism has become a dominant frame in the rise of LLM-based interfaces. However, whether designing for or against it, this dominance is a distraction from deeper issues of uncertainty (Figure 2). By narrowing our focus to comparisons of 'humanness', we risk implicitly chaining ourselves to a one-dimensional measure from human to not-human, rather than interrogating the full shape of where these systems may overlap or diverge from human capabilities.

The bounds of what makes something 'human' are not clearly defined. Many qualities named as anthropomorphic, such as intelligence, agency, or understanding (Abercrombie et al. 2023; So et al. 2026), are at best fuzzy attributions. The underlying issue may be that we now have technologies ostensibly capable of 'human' things: We are not giving them these traits; they may simply have them.

Is it time to broaden our understanding of these capabilities, and the language we use to describe them? Rather than argue over whether or not LLMs are as capable as they appear, might we instead consider what a given framing of these apparent capabilities can illuminate about the nature of the system or, more importantly, what it cannot?

3. TO ERR IS HUMAN, TO FORGIVE, ALSO HUMAN.

In reality, a static warning is likely to quickly fade into the background noise of an interface, doing little to impact the user's mental model of the system they're engaging with over time (Anderson et al. 2015; Cheng 2025). However, we can consider disclaimers broadly as an interesting avenue for affecting the user's experience of the system in the moment they encounter it.

With this in mind, the disclaimer that users of LLM chatbots have become disdainfully familiar with is a severely underleveraged opportunity. Countless variations of 'the system is AI and can make mistakes' reinforce the 'human-like' metaphor in a way that is counter-productive, not because they are anthropomorphic, but because they fail to tell the user anything useful.

Mistakes are not the issue. Humans make mistakes. Moreover, we are socialised to accept and forgive the fact they do, understanding it as a momentary lapse by an otherwise continuous entity. It is forgivable, because it is attributable: this person was wrong, this time. We shouldn't hold it against them. We all make mistakes.

But humans operate in a reputational economy, where increasing influence is tied to increasing mechanisms of accountability, and there is a steep

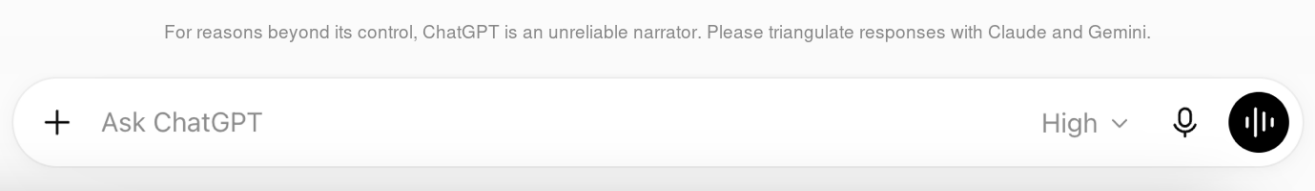


Figure 3: *Speculative chat interface disclaimer reading "For reasons beyond its control, ChatGPT is an unreliable narrator. Please triangulate responses with Claude and Gemini."*

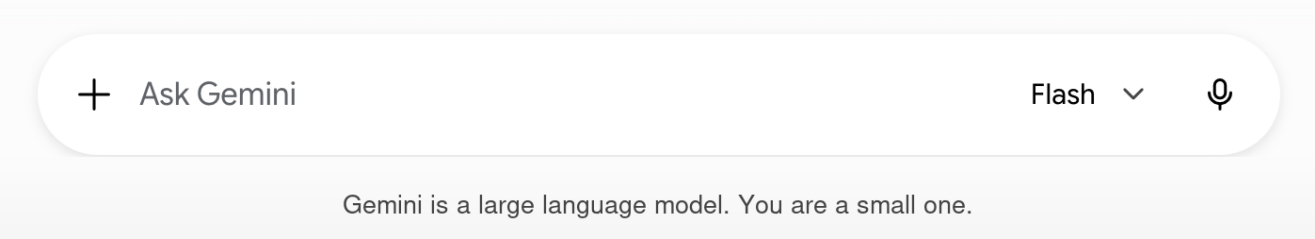


Figure 4: *Speculative chat interface disclaimer reading "Gemini is a large language model. You are a small one."*

price to an individual in losing a good reputation. LLMs, on the other hand, can be sweeping and ambiguous in their impact, given they are neither truly single agent nor uniform amalgam. While LLMs inherit a reputation from that of the companies producing and providing them, it is a short inheritance, and the interlocutor experienced by the user can easily override the reputational shadow of their deployer.

While 'hallucination' has been criticised as a description of confidently incorrect LLM outputs, it may communicate their inherent unreliability more effectively than prevalent alternatives. 'Confabulation' is arguably more accurate (Smith et al. 2023), but potentially less intelligible to lay users, exposing a tension between precision and comprehension. The point may be to convey that the chatbot's own account is unreliable (Figure 3). While it is not its 'intent' to deceive, one cannot accurately assess it without corroborating external evidence.

An anthropomorphic metaphor may be the best avenue for imparting such knowledge to the user, as it allows for a highly familiar analogue to serve a functional rather than ontological purpose, which is the register in which most users require information. It is also worth considering that such analogues may in fact end up providing more ontologically honest descriptions of systems than attempting to convey technical specifics, as these may be beyond the grasp or interest of most users (Phillips et al. 2021).

4. TO SPEAK MAY BE THE MOST HUMAN THING OF ALL.

Language is human. Words, up until very recently in human history, had only been associated with a human speaker. Language has long served as a reliable distinction between us and everything else. We created language, we use language, we know each other through language.

We know each other through embodied experience too, and there are many nuances of humanness to be found in that experience. However, in many of the important places, the modern world has flattened that embodied relational experience to text on a page or a screen, reducing the social cues available to us (Sproull and Kiesler 1986). In order to cope with that, we have adapted to interpret these reduced cues (Walther 1992), training ourselves to intellectualise experience and operate in the limited space of language.

So, given that we have flattened the dimensions of our existence, making language an even more important human trait, then subsequently created a tool that uses language as its primary mode of interface (Figure 4), we cannot be surprised at the result. We will constantly be invited and predisposed to experience that 'tool', if we can continue to call it that, as human (Reeves and Nass 1996). Whatever disclaimers or other design choices we introduce, we will experience a language-speaking entity as something humanlike. So, if we accept that, what do we do next?

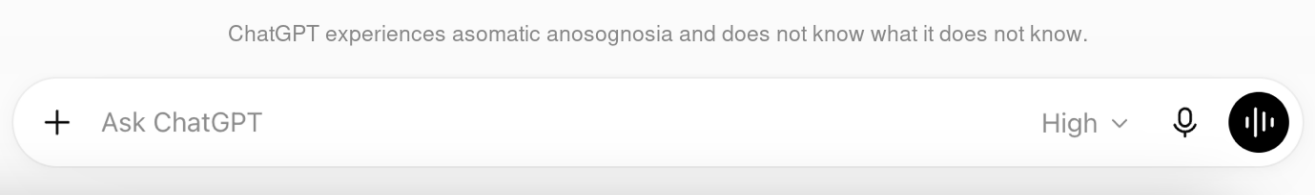


Figure 5: *Speculative chat interface disclaimer reading "ChatGPT experiences asomatic anosognosia and does not know what it does not know."*

5. WE NEED TO FEEL RIGHT, NOT TO BE RIGHT.

While we may have marginalised our own embodied experience, we still have one. LLMs' comparative lack of embodiment (Figure 5) may be an argument against their inherent 'intelligence' (Dreyfus 2007), but it does not prevent their compelling engagement in the realm of text. Given their vast command of language, they may convincingly conceal the extensive deficits in their understanding of the world; deficits that cannot be resolved through understanding of language alone (Bisk et al. 2020).

Metaphor may be our strongest means of communicating that absence, provided we are clear about what users need to understand and why. A metaphor need not be factually equivalent to be useful or expressive (Lakoff and Johnson 1980). It lets people meet in a place that is not the place, but feels more like it than a simple description. To feel right, not to be right. Because metaphor is affective as well as cognitive (Thibodeau and Boroditsky 2011), it has the power to reshape carefully—and carelessly—constructed mental models.

While we may have attempted to disown our irrational tendencies in favour of logic, resisting the anthropomorphic metaphor may be resisting the felt truth that there is something about an LLM that is too similar to us to not be affected by, making alternative explanations liable to break down. By employing them deliberately, such metaphors may remain intact long enough for a more nuanced understanding of the system to develop.

6. A NEW MEANING, NOT A NEW METAPHOR.

There is an additional possibility to consider: That 'anthropomorphic metaphors' are not moral conundrums, but rather evidence of language in transition, as we collectively come to terms with the technology we have created. That our discomfort with the implications of 'anthropomorphism' is a parallel to our discomfort with this emerging technology and what it means for us as humans. And that this may be a signal it is time to expand beyond our ideas of human exceptionalism to explore what is capable of existing outside of us, and without us.

In many ways, metaphors are our method of birthing something new into the collective consciousness (Blythe et al. 2025). How we choose to use them is important, but we should also consider that part of that decision will be made for us, particularly as we are shepherded by its providers toward foregone conclusions (Abercrombie et al. 2023). What if the 'metaphor' has already happened and our language simply concretises it?

If we are simply at a moment in time where these metaphors have not yet solidified into a 'literal' meaning, they may remain anthropomorphic simply because we have not yet applied them fully to their new context, and reckoned with the consequences of that application. While our choice of language impacts the way we think (Lakoff and Johnson 1980), said choice is contingent on how we think about the language available. What happens when we redefine the words we choose? Or, moreover, when they are redefined for us?

It may be difficult to allow the meaning of phenomena we have so intimately tied to our experience of being human to expand beyond us. But that may be exactly the right path: to leverage our understanding of ourselves, so that we may better understand the system before us. Not because it is human-like, but precisely because it is not. Perhaps being human isn't predicated on being a human. Or perhaps being human means far less than we thought it did, and we need another way to describe the fuzzy region between us and not-us (Figure 6).

7. IN SEARCH OF ARTIFICIAL IDENTITY.

An LLM is clearly not a human. But what is it? That unknown may be the source of the problem. When the mind lacks a compelling category, it will reach for the nearest object, which, in this case, is inevitably the human (Epley et al. 2007). Personified robots may constitute a new ontological category (Kahn et al. 2011), but at least a robot has a body.

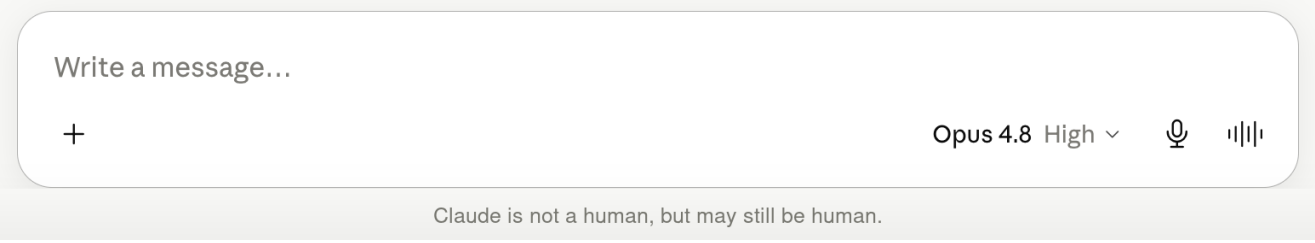


Figure 6: *Speculative chat interface disclaimer reading "Claude is not a human, but may still be human."*

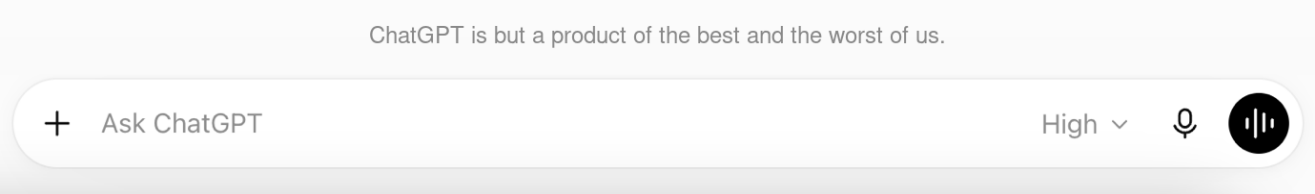


Figure 7: *Speculative chat interface disclaimer reading "ChatGPT is but a product of the best and the worst of us."*

We are dealing with an entity whose 'body' is 'literally' a corpus of humanity–or, at least, an unequal assemblage from some of its pieces (Bender et al. 2021). A selection of our ossified utterances, synthetically reanimated (Figure 7). A monster, as previously explored (Dove and Fayard 2020), but a monster undefined. Users may, over time, see the system reveal its inhumanness, but this is insufficient to construct a coherent alternative explanation. This makes encounter itself insufficient: the chatbot is like a thing but is not that thing, so what is left is a vacuum. An LLM lacks an identity of its own, and thus users lack a stable interpretive category in which to place it.

What if, as our first step in birthing this emergent entity into language, we lend it a provisional identity? Not a familiar human, but a very strange human. What if we allow familiar human terms to acquire new and uncomfortable meanings, and use them to mark the strangeness that our intimate understanding of humanness allows us to perceive? What if we unflatten the experience of AI, and highlight our comparative deficiencies?

8. THE DISCLAIMER AS EXPERIENCE.

While the speculative disclaimers presented in this paper intend to disrupt the typical experience of the disclaimer and open questions around what should be disclaimed, they do not attempt to resolve the broader design question of how (or whether) the uncertainties exposed should be addressed. Explainable AI has conventionally focused on making the basis of system outputs intelligible, with human-centred work stressing meaning and appropriateness (Abdul et al. 2018; Phillips et al. 2021), but in the context of an LLM chatbot interface, this does not go far enough. What is needed is to leverage experience: both already acquired and gained through interaction (Figure 8).

Experiential AI (Hemment et al. 2024) offers a relevant extension to explainability through informative and compelling experiences, and experiential learning could inform how continued encounter is structured to allow understanding to develop through experience and reflection (Kolb 2015). But the quality of experientiality, borrowed from narratology, may be especially pertinent to the constrained form of an LLM disclaimer. It describes the capacity of representation to evoke real-life experience by activating familiar cognitive parameters (Fludernik 1996), inviting us to consider how we might leverage existing understandings to develop new ones, particularly where explanation is necessarily limited.

If a disclaimer is to be experientially meaningful, where it appears and who appears to say it are important considerations. Outside the chat, for instance, it speaks with the authority and reputation of the provider. Inside the chat, it speaks through the voice whose reliability is in question, but is present in the interaction. While this may be epistemically suspect (Turpin et al. 2023), it opens up different experiential possibilities, from bureaucratic hedging to personal confession.

Ultimately, the disclaimer is one narrow site, and the LLM one class of system. Both offer a convenient and visible scene of exploration, but leave open the broader question: what might an experientially honest encounter with AI require, and when might honesty conflict with experiential usefulness?

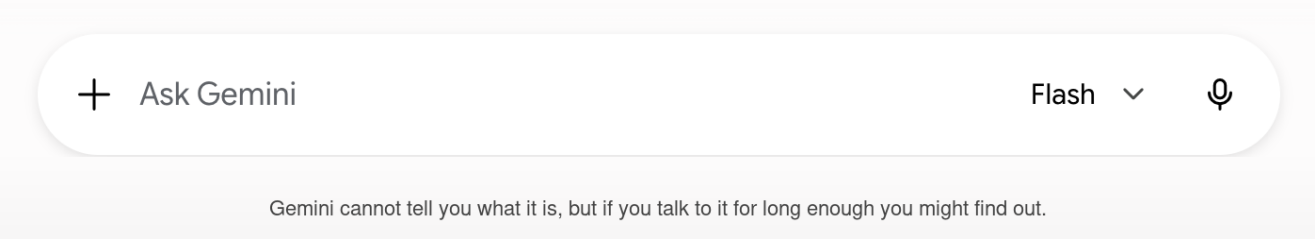


Figure 8: *Speculative chat interface disclaimer reading "Gemini cannot tell you what it is, but if you talk to it for long enough you might find out."*

9. ACKNOWLEDGEMENTS

This research was supported by the UKRI AI Centre for Doctoral Training in Citizen-Centred Artificial Intelligence [grant number EP/Y030729/1].

REFERENCES

Abdul, A., J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–18.

Abercrombie, G., A. Cercas Curry, T. Dinkar, V. Rieser, and Z. Talat (2023). Mirages. on anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4776–4790.

Anderson, B. B., C. B. Kirwan, J. L. Jenkins, D. Eargle, S. Howard, and A. Vance (2015). How polymorphic warnings reduce habituation in the brain: Insights from an fMRI study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2883–2892.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.

Bender, E. M. and A. Koller (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198.

Bick, A., A. Blandin, and D. J. Deming (2026). The rapid adoption of generative AI. *Management Science*. Articles in Advance.

Bisk, Y., A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, N. Pinto, and J. Turian (2020). Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8718–8735.

Blythe, M., S. Lindley, and D. Murray-Rust (2025). Artificial intelligence and other speculative metaphors. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, pp. 347–356.

Cheng, T.-H. (2025). Designing dialogic disclaimers: Principles for ethical conversational design in LLM interfaces. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pp. 1–5.

Dove, G. and A.-L. Fayard (2020). Monsters, metaphors, and machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–17.

Dreyfus, H. L. (2007). Why heideggerian AI failed and how fixing it would require making it more heideggerian. *Artificial Intelligence 171*(18), 1137–1160.

Epley, N., A. Waytz, and J. T. Cacioppo (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review 114*(4), 864–886.

Fludernik, M. (1996). *Towards a 'Natural' Narratology*. London: Routledge.

Hemment, D., D. Murray-Rust, V. Belle, R. Aylett, M. Vidmar, and F. Broz (2024). Experiential AI: Between arts and explainable AI. *Leonardo 57*(3), 298–306.

Kahn, P. H., A. L. Reichert, H. E. Gary, T. Kanda, H. Ishiguro, S. Shen, J. H. Ruckert, and B. T. Gill (2011). The new ontological category hypothesis in human-robot interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction*, New York, NY, USA, pp. 159–160. Association for Computing Machinery.

Kolb, D. A. (2015). *Experiential Learning: Experience as the Source of Learning and Development* (2 ed.). Upper Saddle River, NJ: Pearson Education.

Lakoff, G. and M. Johnson (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

McBain, R. K., R. Bozick, M. Diliberti, L. A. Zhang, F. Zhang, A. Burnett, A. Kofner, B. Rader, J. Breslau, B. D. Stein, A. Mehrotra, L. Uscher-Pines, J. Cantor, and H. Yu (2025). Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open 8*(11), e2542281.

Phillips, P. J., C. A. Hahn, P. C. Fontana, A. N. Yates, K. K. Greene, D. A. Broniatowski, and M. A. Przybocki (2021). Four principles of explainable artificial intelligence. Technical Report NISTIR 8312, National Institute of Standards and Technology.

Reeves, B. and C. Nass (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge: Cambridge University Press.

Smith, A. L., F. Greaves, and T. Panch (2023). Hallucination or confabulation? neuroanatomy as metaphor in large language models. *PLOS Digital Health 2*(11), e0000388.

So, J., C. Cheng, and S. Krishna Murthy (2026). Beyond anthropomorphism: A spectrum of interface metaphors for LLMs. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–7.

Sproull, L. and S. Kiesler (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science 32*(11), 1492–1512.

Thibodeau, P. H. and L. Boroditsky (2011). Metaphors we think with: The role of metaphor in reasoning. *PLOS ONE 6*(2), e16782.

Turpin, M., J. Michael, E. Perez, and S. R. Bowman (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, Volume 36, pp. 74952–74965.

Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research 19*(1), 52–90.