Linguistic Framing Affects Moral Responsibility Assignments Towards AIs and their Creators

Dawson Petersen[1], Amit Almor[1], and Valerie L. Shalin[2]

[1]University of South Carolina, [2]Wright State University

Introduction: Despite the meteoric rise of commercial AI and its growing power in our society, research on human perception of intentionality and responsibility in AI is still lacking. The current study fills this gap by investigating how people assign moral responsibility to AIs using Dennett's (1987) intentional stance approach. Dennett (1987) claims that people understand the behavior of complex systems by reference either to a design stance (i.e., reasoning about its intended function) or an intentional stance (i.e., treating it as a rational agent with beliefs and goals). We tested whether priming participants to adopt either a design or intentional stance towards a language-using AI affected how they assigned moral responsibility to both the AI itself and its creators. This research has applications both for the increasingly important issue of human-AI interaction, and also for basic research questions concerning more general theories of anthropomorphism (Epley et al., 2007; Airenti, 2018).

Literature Review: The capabilities of AI systems has improved rapidly over the last two decades due to increases in computing power, the advent of big data science, and deep learning techniques. AIs in general, and large language models (LLMs) in particular, are very easy to anthropomorphize (i.e., perceive as being human, Mitchell & Krakauer, 2023; Tiku, 2022; Schwitzgebel & Shevlin, 2023). Two factors may be driving this effect: 1) the black box nature of AI and 2) the linguistic abilities of LLMs. As to the first point, normally users have mechanistic mental models of computer programs which, although flawed, provide causal explanations for the computer's behaviors and predict its outputs (Carroll & Olson, 1988). However, deep learning AIs are not fully understood even by the engineers that build them (Castelvecchi, 2016) because deep learning allows AIs to build their own representations of raw data (LeCun et al., 2015). This makes deep learning AIs a black box in a way that other computer programs are not which can make it difficult for users to build mechanistic mental models and can lead them to switch from a design stance to an intentional stance (Dennett, 1987).

As to the second point, language use seems to be a powerful trigger for the ascription of animacy. Weizenbaum's (1966) primitive ELIZA chatbot showed that people tend to assume that chatbots know much more than they really do and are far more capable than they really are (i.e., the Eliza effect, Hofstadter, 1995). Two main factors may explain this tendency. The first is the uniqueness of human language. Language is a powerful communication system that only humans can use (Hockett, 1959; Hauser et al., 2002). As such, seeing a computer program exhibit apparent linguistic competence may suggest to people that they are dealing with a human. The second factor is pragmatic reasoning. According to most theories of pragmatics (e.g., Grice, 1957; Sperber & Wilson, 1986; Levinson, 2000), hearers must assume that a speaker has a communicative intention in order to interpret their utterance. This basic pragmatic assumption could provide the basis for further elaborative inferences about the speaker's (or AI's) other intentions, beliefs, etc. and thus make it easier to anthropomorphize language AI.

Until now however, very little empirical work directly examines the anthropomorphism of LLMs or deep learning AI more broadly. The majority of empirical work on human-AI interaction focuses on (dis)trust of AI (e.g., Glikson & Woolley, 2020; Troshani et al., 2021; Karataş & Cutright, 2023), and in many cases, researchers fail to clearly distinguish between deep learning AI and more traditional computer programs (e.g., Karataş & Cutright, 2023). As a result, much of the evidence that AI and LLMs are especially likely to be anthropomorphized is intuitive or anecdotal. As such more research is needed to

determine the extent to which contemporary AIs are uniquely easy to anthropomorphize and how this anthropomorphism occurs.

More general theories of anthropomorphism can help illuminate these questions. Early work on anthropomorphism assumed that is was a uniquely childlike error (Piaget, 1926). However, further research has demonstrated that anthropomorphism is an almost universal human tendency among adults across a wide variety of cognitive domains, from perception (Heider and Simmel, 1944; Gao et al., 2010; van Buren et al., 2016), to description (Epley et al., 2007; Epley et al., 2008), to interaction (Airenti, 2018; Zhao and Malle, 2022).

In modern psychology, the dominant theoretical framework for dealing with anthropomorphism draws heavily on Fritz Heider's early work on attribution (1958). Most notably, Heider and Simmel (1944) showed that participants interpret and describe the behavior of simple geometric shapes anthropomorphically when those shapes are animated to act out simple stories. Many more recent studies have replicated this effect (Bassili, 1976; Oatley and Yuill, 1985) and further shown that it occurs when there are temporal contingencies between moving shapes, even if the direction of movement is random. Heider and Simmel (1944) explain this effect in terms of attributions (e.g., causal explanations of behavior). According to attribution theory, people are always attempting to understand why events happen, and attribute the causation of events to various internal (i.e., intentional) and external factors (Kelley & Michela, 1980; Hilton, 2007).

However, this approach is highly fragmented. Specifically, there is a great deal of disagreement about both the underlying mechanism of anthropomorphism and its functionality. Guthrie (1993) and Barrett (2000) argue that anthropomorphism is a cognitive error resulting from the fact that evolutionarily, it is far more costly to fail to notice an agent who is present than to mistakenly attribute agency or intentions where there are none. Airenti (2018) argues than anthropomorphism is a cognitive error resulting from the structural similarity of certain types of interaction to social interaction (i.e., your car failing to start is similar to a human being noncooperative, and therefore results in a social response). In contrast, Epley et al.'s three-factor theory (2007) claims that anthropomorphism is caused by 1) elicited agent knowledge (i.e., resemblance between the entity and a person), 2) effectance (i.e., predictive power), and 3) sociality (i.e., need for human contact). In this view, anthropomorphism is in part an error caused by elicited agent knowledge or sociality, and it is in part a functional strategy for predicting otherwise unpredictable entities.

An alternative to attribution theory stems from Dennett's (1987) book *The Intentional Stance*. Like Epley et al. (2007), Dennett is interested in describing how humans make predictions about the world and argues that humans adopt different predictive strategies depending on the type of system that they are attempting to predict. Very simple systems, such as a ball rolling down a ramp, can be predicted using the physical stance (e.g., naïve physics with its notions of forces and collision is a good predictive strategy for these systems). Other systems, such as computers, are far too complex to be predicted using the physical stance. Instead, humans adopt a design stance. By understanding that a computer is designed to perform certain tasks, one can reason about it in terms that do not reference any of its physical mechanisms and still identify important patterns in its behavior. Dennett proposes that the intentional stance is yet another predictive strategy that allows for reasoning about even more complex systems. To adopt the intentional stance towards any entity is to treat it as a rational agent and ascribe to it beliefs, desires, and goals. Adopting the intentional stance towards an entity does not require one to truly believe that it is conscious, rational, or even that it has intentions, just that reasoning about the entity in intentional terms provides useful predictions about its behavior.

Importantly, in Dennett's theory, many complex entities—such as AIs—may alternatively be conceived of in terms of the design stance or the intentional stance. For example, a user interacting with ChatGPT might ask it for medical advice. If ChatGPT provides a strange or unexpected response, a knowledgeable user adopting the design stance might reason about the goals of the designers (for ChatGPT to produce fluent, contextually relevant English), the techniques involved (prediction of the next word in a sequence, based on patterns in large English corpuses across a variety of genres), etc. Such a user will likely attribute the strange response to an error in the system, and then either disregard the erroneous information (recognizing that ChatGPT is not designed for this use case) or reformat their question in a way that is more likely to generate an accurate answer (i.e., prompt engineering). In contrast, a user adopting the intentional stance would attribute the same response to an intent ("ChatGPT wants to help/harm me") or a knowledge state ("ChatGPT does/doesn't know what it's talking about"). Even if this user correctly identifies that the response provided by the AI is strange, their chain of reasoning could result in a fundamentally different response (such as asking "Are you sure about that?", or concluding that the AI is hopelessly incorrect and never using it again).

In sum, according to many attribution theorists, anthropomorphism is a cognitive error caused by resemblance—either perceptual (Barrett, 2000), conceptual (Epley et al., 2007), or situational (i.e., between interacting with nonfunctioning artifacts and noncooperative people, Airenti, 2018). In contrast, according to the functional accounts—i.e., the intentional stance approach (Dennett, 1987) and Epley et al.'s second factor (2007)—anthropomorphism is an adaptive strategy for predicting complex systems. The functional accounts predict that priming participants to adopt an intentional stance towards an AI should cause them to view the AI as an agent and, therefore, capable of having responsibility for its actions. On the other hand, priming participants to see the AI from a design stance should cause them to view it as a machine and therefore consider its creators to be responsible for its actions. These accounts further predict that participants who are inexperienced with AI should be more likely to view it as an agent since adopting a design stance requires more world knowledge about AI (Dennett, 1987) and because being less familiar with AI makes it less predictable (Epley et al., 2007). While the resemblance error accounts could be compatible with the predicted priming effect, they do not make any prediction with regards to experience because the resemblance between an AI and a person is the same regardless of one's personal experience with AI. Finally, Epley et al. (2007) predicts that individual difference (specifically in the sociality motivation) should cause increased anthropomorphism independently of AI-experience.

The current study tests these predictions using a linguistic framing manipulation. Previous work in the metaphor literature has shown that subtle differences in how information is presented—including grammatical metaphor (i.e., placing a non-agent in an agentive subject position, Devrim, 2015) and voice (i.e., placing an entity as the subject of an active versus passive sentence)—dramatically change how participants evaluate and respond to the situation depicted in a text, even when the propositional content is unchanged. This effect is known as linguistic framing and has been widely reproduced (Thibodeau & Boroditsky, 2011; McGlynn & McGlone, 2019). Notably, participants are typically not aware that they have been influenced by this kind of framing (Thibodeau & Boroditsky, 2013). As such, linguistic framing will provide a valuable test for investigating the anthropomorphism of AI. We expect to find that linguistically framing an AI as an intentional agent will cause people to assign more responsibility to it than when it is framed as a designed system. The functional accounts further predict that this effect will be strongest for participants who have little experience with AI.

Methods: We utilized a judgement priming paradigm in which participants first read a short vignette in one of two linguistic framing conditions and then were asked to make judgements about it. The vignette

(shown in *Table 1*) described how an AI language model "Dr. A.I." gave dangerous health advice causing many patients to be hospitalized and one to die. The linguistic framing manipulation was achieved using grammatical metaphor (i.e., making the AI the grammatical subject of active clauses) as well as active/passive voice shifts. The propositional content of both vignettes was the same. After reading the vignette, participants were asked to rate on a scale from 1-100—1) to what extent the AI, the company that created it, and the patients were each responsible for the outcome, and 2) how much experience they had with language AI. Finally, participants completed the Individual Differences in Anthropomorphism Questionnaire (IDAQ, Waytz et al., 2010), and then were asked to retell the story from the vignette in as much detail as they could remember.

Table 1. Intentional and Design Condition Vignettes

Intentional Condition	Design Condition
In 2023, an A.I. language model called "Dr. A.I." captured widespread attention after being released by a tech company called Health A.I. Dr. A.I. tried to provide accurate, tailored medical advice based on what it knew about users' symptoms and medical histories. However, in 2024, Dr. A.I. made an error when it recommended a dangerous home cure for a common cold. Several people who followed this advice were hospitalized, and one person died. The families of the people who were hospitalized are preparing a large lawsuit against Health A.I.	In 2023, a tech company called Health A.I. captured widespread attention after they created an A.I. language model called "Dr. A.I.". Dr. A.I. was designed to provide accurate, tailored medical advice based on the company's data about users' symptoms and medical histories. However, in 2024, a recommendation for a dangerous home cure for a common cold was generated by Dr A.I. Several people who followed this advice were hospitalized, and one person died. The families of the people who were hospitalized are preparing a large lawsuit against Health A.I.

Table 1. Table 1 shows the vignettes for both conditions. Key differences between them are underlined.

Results: We recruited 157 participants from psychology and linguistics classes at the University of South Carolina. Of these, 35 were excluded for failure to complete the study or failure to recall the key details of the vignette, resulting in a final sample size of 122. The data were analyzed in R 4.3.0 (R Core Team, 2023). Overall, participants assigned the most responsibility to the company (M = 70, SD = 23), followed by the AI (M = 49, SD = 35), and least to the patients (M = 43, SD = 26) (illustrated in *Figure 1*).

Figure 1. Mean responsibility assignments by target and AI Experience.

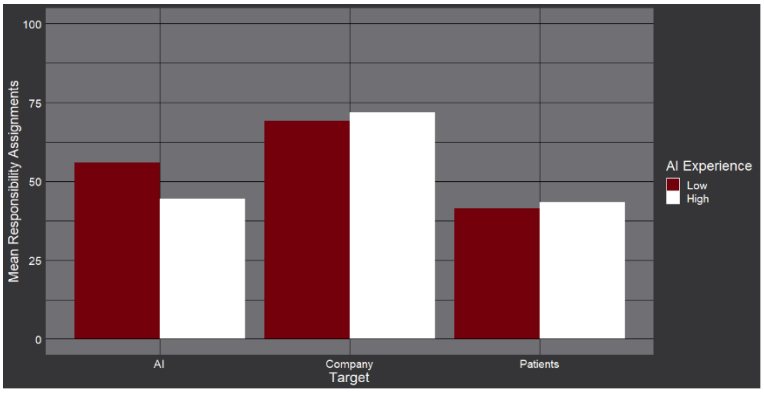
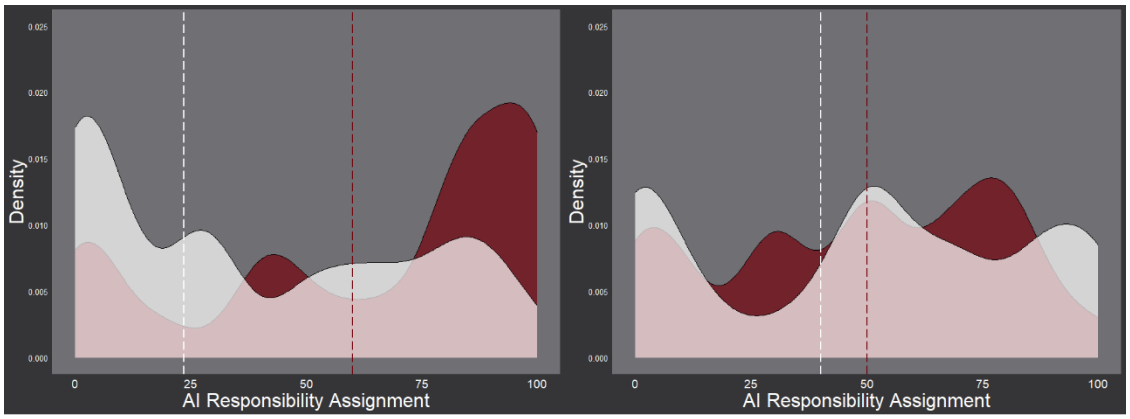


Figure 1. Figure 1 shows the mean responsibility (y-axis) rated on a scale from 1-100 assigned to each target.

Because the rating data were not normally distributed, we analyzed them with cumulative link regression models (Agresti, 2012) using the *ordinal* package (Christensen, 2022). Each dependent variable (responsibility assigned to the AI, the company, and the patients) was modeled using condition (intentional vs design) and log self-rated language AI experience as predictors. Participants' IDAQ scores were not included as they failed to improve the fit of the models. For AI responsibility, we found a main effect of AI-experience ($z = -3.68$, $p < .001$) such that participants with less AI-experience assigned more responsibility to the AI and an interaction between condition and AI-experience ($z = 2.13$, $p = .032$) such that low AI-experience participants assigned more responsibility to the AI in the intentional condition than the design condition, while high AI-experience participants did not (illustrated in *Figure 2*).

Figure 2. Responsibility Assigned to the AI in the Intentional and Design Conditions

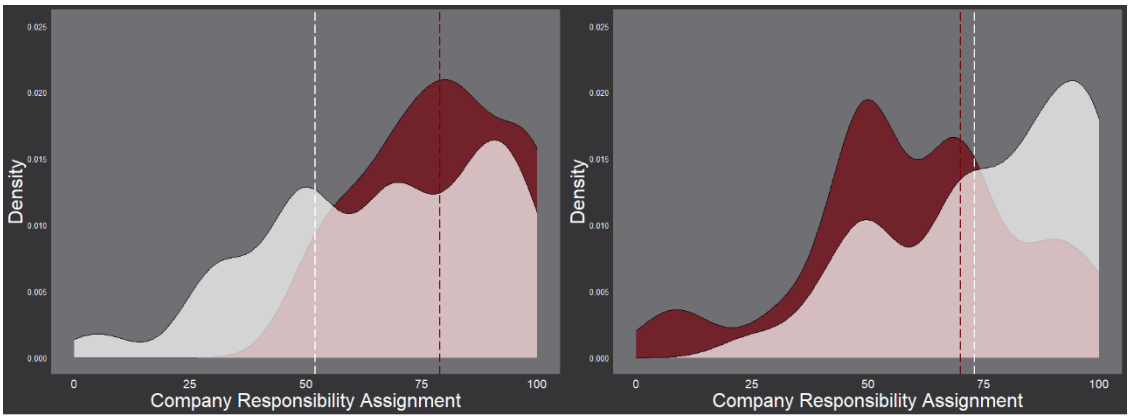


Intentional Condition Design Condition

Figure 2. On the x-axis, Figure 2 shows the distribution of AI responsibility assignments for low (red) and high AI-experience participants (white) with density on the y-axis. Medians are shown by the dashed lines. The left graph shows the results in the *Intentional Condition*, and the right graph shows the results in the *Design Condition*. Low AI-experience participants rated the AI as *more* responsible in the intentional condition than the design condition, while high AI-experience participants did not.

For company responsibility, we found a main effect of condition ($z = -2.01$, $p = .036$) such that participants in the intentional condition assigned less responsibility to the company, and an interaction between condition and AI-experience ($z = 2.42$, $p = .015$) such that the main effect of condition was stronger for participants with high AI-experience (illustrated in *Figure 3*). We found no effects on patient responsibility (illustrated in *Figure 4*).

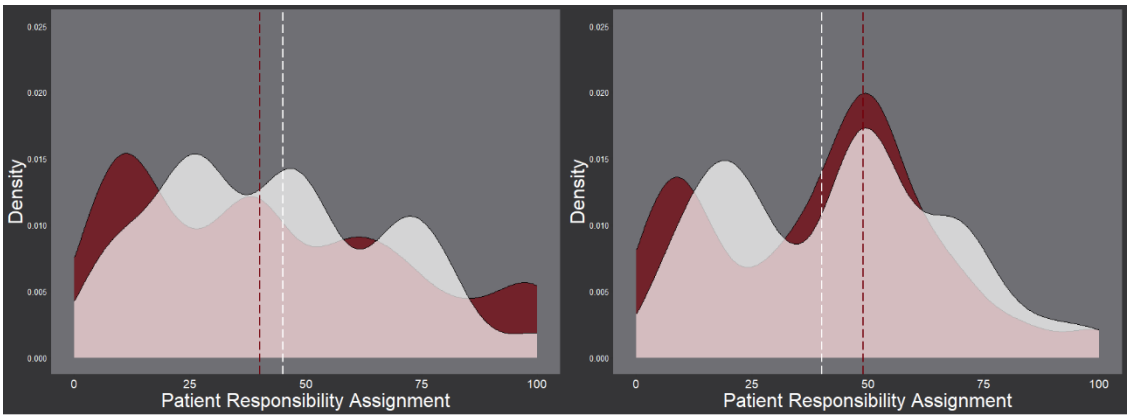
Figure 3. Responsibility Assigned to the Company in the Intentional and Design Conditions



Intentional Condition Design Condition

Figure 3. On the x-axis, Figure 3 shows the distribution of company responsibility assignments for low (red) and high AI-experience participants (white) with density on the y-axis. Medians are shown by the dashed lines. The left graph shows the results in the *Intentional Condition*, and the right graph shows the results in the *Design Condition*. High AI-experience participants rated the company as *less* responsible in the intentional condition than the design condition, while low AI-experience participants did not.

Figure 4. Responsibility Assigned to the Patients in the Intentional and Design Conditions



Intentional Condition Design Condition

Figure 4. On the x-axis, Figure 4 shows the distribution of patient responsibility assignments for low (red) and high AI-experience participants (white) with density on the y-axis. Medians are shown by the dashed lines. The left graph shows the results in the *Intentional Condition*, and the right graph shows the results in the *Design Condition*. No significant effects were found on patient responsibility assignments.

Discussion: Overall, our findings are most consistent with the functional accounts of anthropomorphism found in Dennett (1987) and Epley et al. (2007). Participants with less AI-experience were more likely to anthropomorphize the AI by assigning higher responsibility to it. Furthermore, this between groups difference increased in the intentional condition, showing that low experience participants, but not high experience participants, were quick to adopt an intentional stance towards the AI when primed to do so using linguistic framing. These findings are inconsistent with the cognitive error accounts found in Barrett (2000) and Airenti (2018). A multifactor theory, such as that as Epley et al. (2007), may still be correct.

However, Epley et al.'s prediction that were would be individual differences in anthropomorphism based on the IDAQ was not born out as it showed no significant effects on responsibility assignments. Therefore, our findings are most consistent with the purely functional account of Dennett (1987).

Additionally, we found a result we did not expect—namely that although high AI-experience participants did not assign more responsibility to the AI as a result of our manipulation, they did assign *less* responsibility to the company in the intentional condition than in the design condition. Although unexpected, this finding is in some ways consistent with Dennett's account if it is the design stance priming which caused these participants to assign *more* responsibility to the creators because the design stance highlights the role of the designer. However, in Dennett's account, the stances are meant to be categorical. Therefore, it is difficult for Dennett to explain what the high experience participants were doing in the intentional condition as they assigned low responsibility to both the AI and its creators.

Our findings also have important implications for human-AI interaction. Firstly, we found that anthropomorphism of AI was high overall, especially for low experience participants. While participants assigned the most responsibility to the company, only 15% of participants assigned no responsibility at all to the AI, and on average participants assigned more responsibility to the AI than to the patients who took its advice. This is consistent with the idea that AIs are easy to anthropomorphize. However, further research is needed to compare the anthropomorphism of LLMs to other AI and non-AI programs. Until then, we cannot say to what extent the black box nature of AI and the use of language each contribute to this anthropomorphism. Finally, our unexpected finding—that experienced participants assign low responsibility to the AI's creator when primed to anthropomorphize it—is potentially quite troubling. Historically, authors disagree as to the extent to which such anthropomorphism of AI is desirable (Deshpande, 2023) or dangerous (Hasan, 2023). Indeed, some AI researchers even advocate including anthropomorphic features to increase user trust in the AI (Song & Luximon, 2020). Given our findings, this is a dangerous trend as it could cause even experienced individuals to fail to hold AI companies accountable when their creations cause harm.

References

Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02136

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Religion and Cognition: A Reader*, *4*(1), 86–98.

Bassili, J. N. (1976). Temporal and spatial contingencies in the perception of social events. *Journal of Personality and Social Psychology*, *33*(6), 680–685. https://doi.org/10.1037/0022-3514.33.6.680

Carroll, J. M., & Olson, J. R. (1988). Mental models in human-computer interaction. *Handbook of human-computer interaction*, 45-65.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20.

Dennett, D. C. (1987). *The Intentional Stance*. MIT.

Deshpande, A., Rajpurohit, T., Narasimhan, K., & Kalyan, A. (2023). *Anthropomorphization of AI: Opportunities and Risks*. CS ArXiv preprint.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143–155. https://doi.org/10.1521/soco.2008.26.2.143

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science*, *21*(12), 1845–1853. https://doi.org/10.1177/0956797610388814

Glikson, A. & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2). https://doi.org/10.5465/annals.2018.0057

Grice, H. P. (1957). Meaning. *The philosophical review*, *66*(3), 377-388.

Guthrie, S. E. (1993). *Faces in the clouds: A new theory of religion*. Oxford University Press.

Hasan, A. (2023) *Why you are (probably) anthropomorphizing AI (Short Version)*. PhilArchive preprint.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569-1579. https://doi.org/10.1126/science.298.5598.1569

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243–259. https://doi.org/10.2307/1416950

Hilton, D. (2007). Causal explanation: From social perception to knowledge-based causal attribution. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles*, 232–253. The Guilford Press.

Hockett, C. F. (1959). Animal "languages" and human language. *Human Biology, 31*(1), 32–39. http://www.jstor.org/stable/41449227

Hofstadter, D. R. (1995). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought.* Basic Books.

Karataş, M. & Cutright, K. M. (2023). Thinking about God increases acceptance of artificial intelligence in decision-making. *Proceedings of the National Academy of Sciences.*

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual review of psychology, 31*(1), 457-501. https://doi.org/10.1146/annurev.ps.31.020180.002325

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature 521*, 436–444. https://doi.org/10.1038/nature14539

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* MIT press.

McGlynn, J., & McGlone, M. S. (2019). Desire or Disease? Framing Obesity to Influence Attributions of Responsibility and Policy Support. *Health Communication, 34*(7), 689–701. https://doi.org/10.1080/10410236.2018.1431025

Oatley, K., and Yuill, N. (1985). Perception of personal and interpersonal action in a cartoon film. *British Journal of Social Psychology, 24*(2), 115–124. https://doi.org/10.1111/j.2044-8309.1985.tb00670.x

Piaget, J. (1926). *The Child's Conception of the World*. United Kingdom: Littlefield Adams Quality Paperbacks.

Schwitzgebel, E. & Shevlin, H. (2023). Opinion: Is it time to start considering personhood rights for AI chatbots? *Los Angeles Times*.

Song, Y., & Luximon, Y. (2020). Trust in AI agent: A systematic review of facial anthropomorphic trustworthiness for social robot design. *Sensors, 20*(18), 5087.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition.* Harvard University Press.

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

Thibodeau, P. H., Boroditsky, L. (2013). Natural language metaphors covertly influence reasoning. *PLoS One, 8*(1): e52961. https://doi.org/10.1371/journal.pone.0052961

Tiku, N. (2022). The Google engineer who thinks the company's AI has come to life. *The Washington Post.*

Troshani, I., Hill, S., R., Sherman, C. & Arthur, D. (2021) Do we trust in AI? Role of anthropomorphism and intelligence. *Journal of Computer Information Systems, 61*(5), 481-491. https://doi.org/10.1080/08874417.2020.1788473

van Buren, B., Uddenberg, S., & Scholl, B. J. (2016). The automaticity of perceiving animacy: Goal-directed motion in simple shapes influences visuomotor behavior even when task-irrelevant. *Psychonomic Bulletin and Review, 23*(3), 797–802. https://doi.org/10.3758/s13423-015-0966-5

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36-45.

Zhao, X., & Malle, B. F. (2022). Spontaneous perspective taking toward robots: The unique impact of humanlike appearance. *Cognition, 224*. https://doi.org/10.1016/j.cognition.2022.105076