proceedings

MDPI

Proceeding Paper

The Double-Edged Sword of Anthropomorphism in LLMs †

Madeline G. Reinecke [1,2,*], Fransisca Ting [3,4], Julian Savulescu [1,5,‡] and Ilina Singh [1,2,‡]

1 Uehiro Oxford Institute, University of Oxford, Oxford OX1 1PT, UK
2 Department of Psychiatry, University of Oxford, Oxford OX3 7JX, UK
3 Department of Psychology, University of Toronto, Toronto, ON M5S 3G, Canada
4 Department of Psychological and Brain Sciences, Boston University, Boston, MA 02215, USA
5 Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore
* Correspondence: madeline.reinecke@psych.ox.ac.uk
† Presented at the Online Workshop on Adaptive Education: Harnessing AI for Academic Progress, Online, 12 April 2024.
‡ These authors contributed equally to this work.

Abstract: Humans may have evolved to be "hyperactive agency detectors". Upon hearing a rustle in a pile of leaves, it would be safer to assume that an agent, like a lion, hides beneath (even if there may ultimately be nothing there). Can this evolutionary cognitive mechanism—and related mechanisms of anthropomorphism—explain some of people's contemporary experience with using chatbots (e.g., ChatGPT, Gemini)? In this paper, we sketch how such mechanisms may engender the seemingly irresistible anthropomorphism of large language-based chatbots. We then explore the implications of this within the educational context. Specifically, we argue that people's tendency to perceive a "mind in the machine" is a double-edged sword for educational progress: Though anthropomorphism can facilitate motivation and learning, it may also lead students to trust—and potentially over-trust—content generated by chatbots. To be sure, students do seem to recognize that LLM-generated content may, at times, be inaccurate. We argue, however, that the rise of anthropomorphism towards chatbots will only serve to further camouflage these inaccuracies. We close by considering how research can turn towards aiding students in becoming digitally literate—avoiding the pitfalls caused by perceiving agency and humanlike mental states in chatbots.

Keywords: hyperactive agency detection (HAAD); artificial intelligence; education; anthropomorphism

Academic Editors: Juliana Gerard and Muskaan Singh

Published: 26 February 2025

Citation: Reinecke, M.G.; Ting, F.; Savulescu, J.; Singh, I. The Double-Edged Sword of Anthropomorphism in LLMs. *Proceedings* **2025**, *114*, 4. https://doi.org/10.3390/proceedings2025114004

1. Introduction

In 1966, ELIZA became one of the world's first chatbots—a computer program designed to convincingly mimic human conversation [1,2]. Though, technically speaking, ELIZA was a far-cry from the large language-based chatbots of today (e.g., Gemini, ChatGPT), discussion surrounding these kinds of programs even then was remarkably prescient: "ELIZA shows, if nothing else, how easy it is to create and maintain the illusion of understanding, hence perhaps of judgment deserving credibility. A certain danger lurks there" [1]. In this paper, we address this danger, arguing that cognitive mechanisms related to agency detection and anthropomorphism make users of these technologies more likely to fall victim to such "illusions of understanding" within large language models (LLMs). We then consider why these tendencies may be pernicious when directed towards LLMs in an educational context. Though anthropomorphism may have some benefits on educational

outcomes (e.g., retention; [3]), we argue that the rise of anthropomorphism in chatbots will only serve to further camouflage "hallucinations" and inaccuracies [4,5]. This makes anthropomorphism in LLMs serve as a "double-edged sword" for academic progress. In Section 4, we close by considering a potential intervention to mitigate people's agentive and anthropomorphic tendencies towards LLMs.

2. Mechanisms of Anthropomorphism

2.1. Hyperactive Agency Detection

There is no shortage of cases where humans detect "agency" where there, in fact, isn't any. An eroded hill on Mars gives the impression of looking like a face [6]; shapes following a random path give the impression of "chasing," if oriented in a particular way [7]. Indeed, even the *suggestion* that letters or faces could be present in an image (that, in truth, consisted of pure noise) led participants to report seeing these illusory targets over a third of the time [8]. These kinds of effects are even present, to some extent, in human infants [9].

Why are humans so sensitive to agentive cues? One possibility is that, in the ancestral environment in which *Homo sapiens* evolved, it was adaptive to detect—and perhaps over-detect—agency [10,11]. Upon hearing a rustle in the grass, for example, it would be safest to assume that another agent (e.g., a lion) caused the noise. If this turned out to be false, then life would go on as normal. We would face no greater danger by having overestimated agency in this instance. But if we assumed the opposite, that the rustle came from the wind rather than a lion (when there was, in fact, a lion present), then we face grave risk. Though this evolutionary cognitive mechanism, termed "hyperactive agency detection" (HAAD), is most often discussed as a naturalistic explanation for religious belief [12–14], we see it as relevant also for understanding why people may be predisposed towards perceiving agency in LLMs.

Consider a study from Andersen et al. [15] examining hyperactive agency detection in the lab. Using virtual reality, researchers had participants traverse a forest with the aim of "detect[ing other] beings". At the beginning of the study, researchers tricked participants into believing that there was a 5% or 95% chance of coming across another agent (when, in truth, there were never any other agents to be found). Researchers also manipulated the environment's sensory reliability—changing whether the forest had clear conditions (high sensory reliability) or was clouded in mist (low sensory reliability). Though sensory reliability affected the false alarm rate—such that participants "detected" agents more often in the misty forest than the clear forest—what mattered most were their priors. Believing that there was a 95% chance of seeing another being in the forest, in fact, led participants to "see" such beings more often.

These findings deliver two primary insights for thinking about anthropomorphism in LLMs: First, there is a sense in which society currently stands in a technologically misty forest. There is widespread debate surrounding whether LLMs count as tools or as agents [16], whether they demonstrate "sparks of general intelligence" [17], and so on. Unlike a clear forest, where agents can be distinguished from non-agents with ease, people's present experience with rapidly advancing artificial intelligence (AI) may mirror the misty forest, where much remains unclear and in flux. The second insight concerns the importance of people's prior beliefs in detecting agency. When participants in this study *believed* that they had a high chance of encountering another agent, their experience in the forest aligned with these expectations. This may similarly occur in the context of LLMs. For example, former Google engineer Blake Lemoine became convinced that LaMDA, a large language model in development at the time, was sentient and deserved to be treated as a Google employee [18]. Lemoine's convictions may have stemmed from how he interacted with LaMDA, such that his questions (e.g., "I'm generally assuming that you would like

more people at Google to know that you're sentient. Is that true?" [19]) confirmed his pre-existing beliefs about the model and its capacities.

Emerging research suggests that people do anthropomorphize ChatGPT [20], particularly with increased exposure [21]. People may even do so without explicit awareness [22]. This could lead to a concerning feedback loop, such that people's anthropomorphic tendencies influence how they engage with LLMs (which then reinforces their beliefs about LLMs' agentive qualities). This cycle, rooted in confirmation bias [23], would fuel the perception of agency where it likely does not exist—potentially affecting experts and non-experts alike [24,25].

2.2. Language as Agentive Cue

Large language models may invite anthropomorphism for another reason. Mimicking humanlike language and conversation is the cornerstone of what LLMs are designed to accomplish [26]. No longer is language use uniquely human [27]. This advance bolsters the likelihood of perceiving LLMs as agents: Communication, whether simple or complex, makes sense only within an agentive context [28]. Given that language likely evolved to allow agents to "share knowledge, thoughts, and feelings with one another" [29], language serves as arguably one of the most direct and perceptible reflections of others' conscious minds. When we perceive others as capable of communicating in this way, the more likely we are to perceive them as agents with complex mental states [30].

Even preverbal infants make this link, inferring agency by observing whether a given entity can communicate. Across a swath of papers, having the ability to converse emerges as one of the most reliable means by which infants establish agency in others [31–35]. For example, in one study [31], experimenters presented 12–13-month-old infants with a live scene where an actor engaged in a contingent, back-and-forth interaction with a novel box covered in brown fur: The experimenter spoke to the box in human speech, and the box "spoke" back in fluctuating beeps, mimicking variability in natural language [36]. Despite the lack of morphological similarities to familiar agents, infants still treated the communicative box as an agent—following its "gaze" when it turned to the left or right. This kind of gaze following is common when infants interact with other people, but this behavior disappears when babies interact with inanimate objects [31–35]. Infants also rely more heavily on language use and conversational cues when determining agency as compared to other features—such as whether an entity has morphological similarities to other agents [37–39], the capacity to self-propel [40,41], or the capacity to engage in non-conversational contingent interaction with other agents [31]. This literature, taken together, suggests that humans may have an early-emerging, if not innate, tendency to link communicative capacity with agency. Given that contemporary LLMs far surpass this minimal conversational threshold, we anticipate that LLM output may very well trigger agentive inferences in users across the lifespan.

3. Anthropomorphism as a Double-Edged Sword in Education

Having a predisposition towards anthropomorphizing LLMs may be, on its own, value neutral. But this predisposition may open the door to a multitude of downstream harms for users (e.g., related to manipulation, coercion, and privacy [42]). We recognize that these harms may be especially rampant within an educational context, given students' rapid adoption of these technologies as learning aids [43,44]. In the next section, we consider some of the consequences for education—both beneficial and detrimental—implicated in LLM-oriented anthropomorphism.

3.1. Benefits of Anthropomorphism

Though emerging research suggests that students' use of ChatGPT can promote at least some aspects of engagement (e.g., completing tasks on time [45]), there has been little to no research examining the intersection between LLMs, anthropomorphism, and educational outcomes. In light of this gap, we draw on research regarding the effects of anthropomorphism in non-AI-based teaching materials. On the whole, this literature suggests that some amount of anthropomorphic detail in teaching materials may benefit students. In a series of studies examining the relationship between anthropomorphism in teaching illustrations and learning outcomes for primary and secondary school students, for example, Schneider et al. observed that having some amount (but not the maximum amount) of anthropomorphic content in illustrations improved retention and transfer learning. Anthropomorphism also enhanced students' sense of intrinsic motivation [46]. In related work completed within secondary school and university settings, anthropomorphic teaching materials again outperformed non-anthropomorphic and control materials [3]. Though effect sizes vary across studies, a meta-analysis from Brom et al. examining the use of facial anthropomorphism and pleasant colors in teaching materials indicates that these inventions, on the whole, tend to improve academic outcomes (e.g., transfer, comprehension, retention), particularly for younger students [47]. These findings suggest that learners' anthropomorphism of LLMs may be an asset for education, rather than a hindrance.

We note, however, that this literature suggests that even traditional learning materials can be too anthropomorphic. In research from Schneider et al., for example, students who received learning materials with the highest degree of anthropomorphism fared no better than students who received materials with the lowest degree of anthropomorphism. This may draw on students facing extraneous cognitive load in the high anthropomorphism case [46]. Further, in the above studies, all instances of anthropomorphism consisted of decorative additions to original teaching materials (e.g., static images of platelets with arms and legs [46]). Further research is needed to determine whether these benefits extend to LLMs in an educational context. At present, students in the United Kingdom most often use LLMs as "AI tutors", engaging in back-and-forth conversation about course content [43]. This could, in principle, solve the famed "2 sigma" problem [48], where students with access to a personal tutor have been shown to outperform students taught in a traditional classroom setting by up to two standard deviations. Indeed, some researchers have already begun to implement and evaluate such LLM-based tutoring strategies, yielding promising results thus far [49,50]. Whether *anthropomorphism* in this context gives rise to beneficial educational outcomes, however, remains to be seen.

3.2. Drawbacks of Anthropomorphism

We also recognize that there is a range of educational disadvantages that may emerge from anthropomorphizing chatbots. We focus here on one chief concern: These systems remain prone to "hallucination," meaning that they generate inaccurate statements that are endorsed by the model as correct [51]. Though this technical challenge may dissipate (or even disappear) with time, it continues to be an issue for the most cutting-edge models presently available. In a line of research testing how often LLMs would generate illusory citations (when specifically prompted to produce real and accurate citations), ChatGPT (powered by GPT-3.5 and GPT-4) hallucinated at least some of the time [5]. At its best, GPT-4's reference list was 11% hallucinated; at its worst, its reference list was 43% hallucinated. GPT-3.5—the freely accessible model from OpenAI—fell even further short. Both of these models also struggled, to some degree, in applying "real citations [that were] appropriate for the answer to [the] question[s]" [5]. This is critical in an educational sphere: If a chatbot makes empirical claims with references, it should present the referenced material accurately.

Yet, GPT-4 provided "proper source[s] for answering [a given] question" only 57–89% of the time, depending on question specificity [5]. Insofar as hallucination remains a problem for LLMs, we see this as deep cause for concern.

This concern stems, in part, from the relationship between anthropomorphism and trust in artificial intelligence (e.g., [52–58]). When users of voice assistants, like Amazon's Alexa, anthropomorphize their assistant ("[Alexa is] fun-loving"), they also tend to trust the assistant ("Although I may not know exactly how Alexa works, I know how to use it to perform tasks I want Alexa to do") and treat their relationship with the assistant as more akin to a human relationship ("I feel guilty when asking too much of Alexa"; [57]). In a similar vein, people tend to trust autonomous vehicles more if the vehicles have been augmented with anthropomorphic features (e.g., having a voice, name, gender; [58]).

We expect these effects to translate to LLMs, such that people who see chatbots as more humanlike will likely trust both accurate and inaccurate AI-generated content more often. Indeed, evidence from Cohn et al. suggests that, after engaging with a pseudo-LLM, participants' anthropomorphic beliefs predicted how trustworthy they found the LLM to be [52]. To be sure, students do report being aware that these systems can generate inaccurate content through hallucination [59]; there are even instances where student users would have done better academically if they had trusted LLM-generated content more [60]. But the problem remains: Students remain generally unaware of how often hallucination occurs [43], meaning that their trust in AI-generated content may be unfounded. We expect this challenge to become even greater as LLMs become further anthropomorphic [42], persuasive [61], and potentially even deceptive [62,63].

4. Reducing the Harm of Anthropomorphism

What can be done to mitigate the harms that students face as a result of anthropomorphizing LLMs, such as being more likely to trust hallucinated content? We speculate that promising interventions in other areas—such as those countering people's acceptance of misinformation, more broadly—may also be successful here.

A growing amount of literature suggests that inoculating the public against misinformation by "preexposing them to severely weakened doses" of false content and then demonstrating their falsity [64]) can diminish susceptibility to misleading information [64–68]. For example, in a set of studies completed in the lab and in the field, researchers found that having people watch videos that deploy and then highlight common manipulation techniques—like using emotional language—actually reduced their willingness to share manipulative content on social media themselves [65]. This "inoculation" technique may dampen people's vulnerability to misinformation across a variety of domains, including climate change [66,68] and vaccination attitudes [67]. Further, these benefits seem to persist beyond the initial intervention (e.g., [67]), suggesting that resistance to misinformation may last over time.

As applied to LLMs, an inoculation intervention might involve "prebunking" (i.e., providing a clear explanation of how LLMs work and, at times, hallucinate), followed by a "microdose" of misinformation from an LLM (that is then revealed to be hallucinated; [64]). For example, an instructor might show students an AI-generated analysis of *Hamlet*, which would likely, for the most part, be accurate. The instructor would then highlight any hallucinations that exist within the AI-generated text, such as an analysis of a scene that never existed. Demonstrating, through examples, how generative AI can convincingly produce true and false statements may help students build "attitudinal immunity" [69]. It may even be helpful to prebunk by describing humans' natural inclination to anthropomorphize (and how this can contribute towards over-trusting AI systems), calling on the literature described in Sections 2 and 3. In research probing misinformation spread surrounding cli-

mate change, for instance, forewarning participants that "politically motivated groups use misleading tactics to try to convince the public that there is a lot of disagreement between scientists" was effective in strengthening their resistance to subsequent climate-related misinformation [66].

Though research is needed to determine whether an inoculation-based approach would be successful in this context, we expect that this kind of strategy may aid students in being digitally literate—by limiting their anthropomorphism through awareness of human psychology, and by having a healthy skepticism of AI-generated content. This approach dovetails with other calls to promote critical thinking and appropriate skepticism in Internet users, more broadly [70,71]. Given that individuals may vary in their propensity to anthropomorphize [72], as well as engage in critical thinking [73], we see this kind of intervention as valuable for investigation across users. Determining how to reduce people's confidence in AI-based misinformation is an essential avenue for future work.

5. Conclusions

In sum, we see the rise of large language models as a remarkable technological achievement, one that comes with incredible promise and need for caution. In particular, we argue that certain cognitive mechanisms related to anthropomorphism—such as hyperactive agency detection and language as an agentive cue—may engender undue trust in chatbots. Further research should examine how users can be supported to calibrate their trust in LLMs, potentially through inoculation-based intervention. We see this as key for harnessing the benefits (and diminishing the drawbacks) of LLMs in the educational sphere and beyond.

Author Contributions: Conceptualization, M.G.R.; writing—original draft preparation, M.G.R; writing—review and editing, M.G.R., F.T., J.S. and I.S.; supervision, J.S. and I.S.; funding acquisition, J.S. and I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in whole, or in part, by the Wellcome Trust [Wellcome Centre for Ethics and Humanities 203132/Z/16/Z], the National Institute for Health and Care Research (NIHR) Oxford Health Biomedical Research Centre [NIHR203316], and the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: -AISG3-GV-2023-012). The views expressed are those of the author(s) and not necessarily those of the Wellcome Trust, NIHR, the Department of Health and Social Care, or the National Research Foundation, Singapore.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: We thank the members of the Centre for Mind and Morality at the University of Toronto for helpful feedback on ideas discussed in this paper, as well as the organizers and attendees of "Adaptive Education: Harnessing AI for Academic Progress".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
2. Shah, H.; Warwick, K.; Vallverdú, J.; Wu, D. Can machines talk? Comparison of ELIZA with modern dialogue systems. *Comput. Hum. Behav.* **2016**, *58*, 278–295. [CrossRef]
3. Schneider, S.; Nebel, S.; Beege, M.; Rey, G.D. Anthropomorphism in decorative pictures: Benefit or harm for learning? *J. Educ. Psychol.* **2018**, *110*, 218. [CrossRef]

4. Athaluri, S.A.; Manthena, S.V.; Kesapragada, V.K.M.; Yarlagadda, V.; Dave, T.; Duddumpudi, R.T.S. Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus* **2023**, *15*, e37432. [CrossRef] [PubMed]
5. Buchanan, J.; Hill, S.; Shapoval, O. ChatGPT hallucinates non-existent citations: Evidence from economics. *Am. Econ.* **2024**, *69*, 80–87. [CrossRef]
6. Carlotto, M.J. Digital imagery analysis of unusual Martian surface features. *Appl. Opt.* **1988**, *27*, 1926–1933. [CrossRef] [PubMed]
7. Gao, T.; Newman, G.E.; Scholl, B.J. The psychophysics of chasing: A case study in the perception of animacy. *Cogn. Psychol.* **2009**, *59*, 154–179. [CrossRef] [PubMed]
8. Liu, J.; Li, J.; Feng, L.; Li, L.; Tian, J.; Lee, K. Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia. *Cortex* **2014**, *53*, 60–77. [CrossRef] [PubMed]
9. Kato, M.; Mugitani, R. Pareidolia in infants. *PLoS ONE* **2015**, *10*, e0118539. [CrossRef]
10. Guthrie, S.E. *Faces in the Clouds: A New Theory of Religion*; Oxford University Press: Oxford, UK, 1995.
11. Barrett, J.L. *Why Would Anyone Believe in God?* AltaMira Press: Walnut Creek, CA, USA, 2004.
12. Barrett, J.L. Exploring the natural foundations of religion. *Trends Cogn. Sci.* **2000**, *4*, 29–34. [CrossRef]
13. Barrett, J.L. The Naturalness of Religious Concepts: An Emerging Cognitive Science of Religion. In *New Approaches to the Study of Religion*; Antes, P., Geertz, A.W., Warne, R.R., Eds.; De Gruyter: Berlin, Germany, 2004.
14. Barrett, J.L.; Lanman, J.A. The science of religious beliefs. *Religion* **2008**, *38*, 109–124. [CrossRef]
15. Andersen, M.; Pfeiffer, T.; Müller, S.; Schjoedt, U. Agency detection in predictive minds: A virtual reality study. *Relig. Brain Behav.* **2019**, *9*, 52–64. [CrossRef]
16. Sedlakova, J.; Trachsel, M. Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent? *Am. J. Bioeth.* **2023**, *23*, 4–13. [CrossRef]
17. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv* **2023**, arXiv:2303.12712.
18. Tiku, N. The Google engineer who thinks the company's AI has come to life. Available online: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/ (accessed on 9 July 2024).
19. Lemoine, B. Is LaMDA Sentient?—An Interview. Available online: https://www.e-flux.com/notes/475146/is-lamda-sentient-an-interview (accessed on 9 July 2024).
20. Colombatto, C.; Fleming, S.M. Folk psychological attributions of consciousness to large language models. *Neurosci. Conscious.* **2024**, *2024*, niae013. [CrossRef]
21. Jacobs, O.; Pazhoohi, F.; Kingstone, A. Brief exposure increases mind perception to ChatGPT and is moderated by the individual propensity to anthropomorphize. Available online: https://osf.io/preprints/psyarxiv/pn29d (accessed on 9 July 2024).
22. Kim, Y.; Sundar, S.S. Anthropomorphism of computers: Is it mindful or mindless? *Comput. Hum. Behav.* **2012**, *28*, 241–250. [CrossRef]
23. Klayman, J. Varieties of confirmation bias. *Psychol. Learn. Motiv.* **1995**, *32*, 385–418.
24. van den Eeden, C.A.; de Poot, C.J.; van Koppen, P.J. The forensic confirmation bias: A comparison between experts and novices. *J. Forensic Sci.* **2019**, *64*, 120–126. [CrossRef] [PubMed]
25. Machery, E. The illusion of expertise. In *Experimental Philosophy, Rationalism, and Naturalism*; Routledge: Abingdon, UK, 2015; pp. 188–203.
26. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Barnes, N.; Mian, A. A comprehensive overview of large language models. *arXiv* **2023**, arXiv:2307.06435.
27. Hauser, M.D.; Chomsky, N.; Fitch, W.T. The faculty of language: What is it, who has it, and how did it evolve? *Science* **2002**, *298*, 1569–1579. [CrossRef]
28. Grice, H.P. Meaning. *Philos. Rev.* **1957**, *66*, 377–388. [CrossRef]
29. Fedorenko, E.; Piantadosi, S.T.; Gibson, E.A. Language is primarily a tool for communication rather than thought. *Nature* **2024**, *630*, 575–586. [CrossRef]
30. Gray, H.M.; Gray, K.; Wegner, D.M. Dimensions of mind perception. *Science* **2007**, *315*, 619. [CrossRef]
31. Beier, J.S.; Carey, S. Contingency is not enough: Social context guides third-party attributions of intentional agency. *Dev. Psychol.* **2014**, *50*, 889. [CrossRef] [PubMed]
32. Johnson, S.; Slaughter, V.; Carey, S. Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Dev. Sci.* **1998**, *1*, 233–238. [CrossRef]
33. Johnson, S.C.; Shimizu, Y.A.; Ok, S.J. Actors and actions: The role of agent behavior in infants' attribution of goals. *Cogn. Dev.* **2007**, *22*, 310–322. [CrossRef] [PubMed]
34. Johnson, S.C.; Bolz, M.; Carter, E.; Mandsanger, J.; Teichner, A.; Zettler, P. Calculating the attentional orientation of an unfamiliar agent in infancy. *Cogn. Dev.* **2008**, *23*, 24–37. [CrossRef]
35. Johnson, S.C. Detecting agents. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **2003**, *358*, 549–559. [CrossRef]

36. Tauzin, T.; Gergely, G. Variability of signal sequences in turn-taking exchanges induces agency attribution in 10.5-mo-olds. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 15441–15446. [CrossRef]
37. Sloane, S.; Baillargeon, R.; Premack, D. Do infants have a sense of fairness? *Psychol. Sci.* **2012**, *23*, 196–204. [CrossRef]
38. Buyukozer Dawkins, M.; Sloane, S.; Baillargeon, R. Do infants in the first year of life expect equal resource allocations? *Front. Psychol.* **2019**, *10*, 116. [CrossRef]
39. Setoh, P.; Wu, D.; Baillargeon, R.; Gelman, R. Young infants have biological expectations about animals. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 15937–15942. [CrossRef]
40. Csibra, G. Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition* **2008**, *107*, 705–717. [CrossRef] [PubMed]
41. Kamewari, K.; Kato, M.; Kanda, T.; Ishiguro, H.; Hiraki, K. Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion. *Cogn. Dev.* **2005**, *20*, 303–320. [CrossRef]
42. Gabriel, I.; Manzini, A.; Keeling, G.; Hendricks, L.A.; Rieser, V.; Iqbal, H.; Tomašev, N.; Ktena, I.; Kenton, Z.; Rodriguez, M.; et al. The ethics of advanced AI assistants. *arXiv* **2024**, arXiv:2404.16244.
43. Freeman, J. *Provide or Punish? Students' Views on Generative AI in Higher Education*; Higher Education Policy Institute: Oxford, UK, 2024.
44. Von Garrel, J.; Mayer, J. Artificial Intelligence in studies—Use of ChatGPT and AI-based tools among students in Germany. *Humanit. Soc. Sci. Commun.* **2023**, *10*. [CrossRef]
45. Lo, C.K.; Hew, K.F.; Jong, M.S.y. The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Comput. Educ.* **2024**, *219*, 105100. [CrossRef]
46. Schneider, S.; Häßler, A.; Habermeyer, T.; Beege, M.; Rey, G.D. The more human, the higher the performance? Examining the effects of anthropomorphism on learning with media. *J. Educ. Psychol.* **2019**, *111*, 57. [CrossRef]
47. Brom, C.; Starkova, T.; D'Mello, S.K. How effective is emotional design? A meta-analysis on facial anthropomorphisms and pleasant colors during multimedia learning. *Educ. Res. Rev.* **2018**, *25*, 100–119. [CrossRef]
48. Bloom, B.S. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educ. Res.* **1984**, *13*, 4–16. [CrossRef]
49. Pal Chowdhury, S.; Zouhar, V.; Sachan, M. AutoTutor meets Large Language Models: A Language Model Tutor with Rich Pedagogy and Guardrails. In Proceedings of the Eleventh ACM Conference on Learning @ Scale, Atlanta, GA, USA, 18–20 July 2024; pp. 5–15.
50. Jiang, Z.; Jiang, M. Beyond Answers: Large Language Model-Powered Tutoring System in Physics Education for Deep Learning and Precise Understanding. *arXiv* **2024**, arXiv:2406.10934.
51. Perković, G.; Drobnjak, A.; Botički, I. Hallucinations in LLMs: Understanding and Addressing Challenges. In Proceedings of the 2024 47th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 20–24 May 2024; pp. 2084–2088.
52. Cohn, M.; Pushkarna, M.; Olanubi, G.O.; Moran, J.M.; Padgett, D.; Mengesha, Z.; Heldreth, C. Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024; pp. 1–15.
53. De Visser, E.J.; Monfort, S.S.; Goodyear, K.; Lu, L.; O'Hara, M.; Lee, M.R.; Parasuraman, R.; Krueger, F. A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents. *Hum. Factors* **2017**, *59*, 116–133. [CrossRef] [PubMed]
54. Kulms, P.; Kopp, S. More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human-agent cooperation. In Proceedings of the Mensch und Computer 2019, Hamburg, Germany, 8–11 September 2019; pp. 31–42.
55. Natarajan, M.; Gombolay, M. Effects of anthropomorphism and accountability on trust in human robot interaction. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, UK, 23–26 March 2020; pp. 33–42.
56. Polyportis, A.; Pahos, N. Understanding students' adoption of the ChatGPT chatbot in higher education: The role of anthropomorphism, trust, design novelty and institutional policy. *Behav. Inf. Technol.* **2024**, *44*, 1–22. [CrossRef]
57. Seymour, W.; Van Kleek, M. Exploring interactions between trust, anthropomorphism, and relationship development in voice assistants. In Proceedings of the ACM on Human-Computer Interaction, Virtual Conference, 8–13 May 2021; pp. 1–16.
58. Waytz, A.; Heafner, J.; Epley, N. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.* **2014**, *52*, 113–117. [CrossRef]
59. Ngo, T.T.A. The perception by university students of the use of ChatGPT in education. *Int. J. Emerg. Technol. Learn. (Online)* **2023**, *18*, 4. [CrossRef]
60. Zhang, P. Taking advice from ChatGPT. *arXiv* **2023**, arXiv:2305.11888.
61. Matz, S.; Teeny, J.; Vaid, S.S.; Peters, H.; Harari, G.; Cerf, M. The potential of generative AI for personalized persuasion at scale. *Sci. Rep.* **2024**, *14*, 4692. [CrossRef] [PubMed]
62. Hagendorff, T. Deception abilities emerged in large language models. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2317967121. [CrossRef] [PubMed]

63. Leong, B.; Selinger, E. Robot eyes wide shut: Understanding dishonest anthropomorphism. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 299–308.
64. Traberg, C.S.; Roozenbeek, J.; van der Linden, S. Psychological inoculation against misinformation: Current evidence and future directions. *ANNALS Am. Acad. Political Soc. Sci.* **2022**, *700*, 136–151. [CrossRef]
65. Roozenbeek, J.; Van Der Linden, S.; Goldberg, B.; Rathje, S.; Lewandowsky, S. Psychological inoculation improves resilience against misinformation on social media. *Sci. Adv.* **2022**, *8*, eabo6254. [CrossRef]
66. Van der Linden, S.; Leiserowitz, A.; Rosenthal, S.; Maibach, E. Inoculating the public against misinformation about climate change. *Glob. Chall.* **2017**, *1*, 1600008. [CrossRef] [PubMed]
67. Van der Linden, S.; Dixon, G.; Clarke, C.; Cook, J. Inoculating against COVID-19 vaccine misinformation. *EClinicalMedicine* **2021**, *33*, 100772. [CrossRef]
68. Maertens, R.; Anseel, F.; van der Linden, S. Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects. *J. Environ. Psychol.* **2020**, *70*, 101455. [CrossRef]
69. Lewandowsky, S.; Van Der Linden, S. Countering misinformation and fake news through inoculation and prebunking. *Eur. Rev. Soc. Psychol.* **2021**, *32*, 348–384. [CrossRef]
70. Graham, L.; Metaxas, P.T. "Of course it's true; I saw it on the Internet!" critical thinking in the Internet era. *Commun. ACM* **2003**, *46*, 70–75. [CrossRef]
71. Browne, M.N.; Freeman, K.E.; Williamson, C.L. The importance of critical thinking for student use of the Internet. *Coll. Stud. J.* **2000**, *34*, 391–398.
72. Waytz, A.; Cacioppo, J.; Epley, N. Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspect. Psychol. Sci.* **2010**, *5*, 219–232. [CrossRef]
73. Campitelli, G.; Labollita, M. Correlations of cognitive reflection with judgments and choices. *Judgm. Decis. Mak.* **2010**, *5*, 182–191. [CrossRef]