The Cognitive Biases Behind Inflationary and Deflationary Claims about Large Language Models

Stefano Palminteri(1, 2,*), Giada Pistilli(3)

(1) Laboratoire des Neurosciences Cognitives et Computationnelles, Institut National de la Santé et de la Recherche Médicale, Paris, France

(2) Departement d'Etudes Cognitives, Ecole Normale Supérieure, Paris, France.

(3) Hugging Face, Paris, France

(*) Mail to: stefano.palminteri@ens.fr

Abstract

The rapid rise of Large Language Models (LLMs) has sparked intense debate across multiple academic disciplines. While some argue that LLMs represent a significant step toward artificial general intelligence (AGI) or even machine consciousness (inflationary claims), others dismiss them as mere trickster artifacts lacking genuine cognitive abilities (deflationary claims). We argue that both extremes may be shaped or exacerbated by common cognitive biases, including cognitive dissonance, wishful thinking, and the illusion of depth of understanding, which distort reality to our own advantage. By showcasing how these distortions may easily emerge in both scientific and public discourse, we advocate for a measured approach— skeptical open mind - that recognizes the cognitive abilities of LLMs as worthy of scientific investigation while remaining conservative concerning exaggerated claims regarding their cognitive status.

Acknowledgements

SP acknowledges support from the European Research Council consolidator grant (RaReMem: 101043804) three Agence Nationale de la Recherche grants (CogFinAgent: ANR-21-CE23-0002-02; RELATIVE: ANR-21-CE37- 750 0008-01; RANGE: ANR-21-CE28-0024-01), the Alexander Von Humbolt foundation and a Google unrestricted gift.

Introduction

The question of whether Large Language Models (LLMs) represent a scientific revolution remains open. While it is difficult to recognize a paradigm shift from within, the sheer scale of disruption caused by LLMs across artificial intelligence and related fields suggests that they may represent a major turning point (Sejnowski, 2023, 2024). Unlike previous trends (expert systems, reinforcement learning, early neural networks), LLMs have drawn unprecedented interdisciplinary engagement, spanning philosophy, cognitive psychology, computer science, and even political, education, economic sciences, and medicine (Demszky et al., 2023; Kasneci et al., 2023; Li et al., 2023, 2023; Yan et al., 2024).

In this multidisciplinary context, the extraordinary performance and the rapid rise of LLMs have fueled polarized discourse. On the one hand, their ability to generate human-like language, solve complex problems, pass legal and medical exams, and even learn new skills on the fly (in context learning), has led some to proclaim the imminent arrival of Artificial General Inteligence (AGI)[1], the singularity or even conscious machines (inflationary claims) (Bubeck et al., 2023; Highmore, 2024). On the other hand, skeptics argue that LLMs are merely "stochastic parrots"—statistical models with no real intelligence, let alone consciousness (deflationary claims) (Bender et al., 2021). The high stakes of this debate—from ethical concerns to policy implications—exacerbate polarization, making it imperative and urgent to critically assess which cognitive processes or biases may contribute to these divergent views (Metzinger, 2021).

Both inflationary and deflationary stances are neither inherently good nor bad unless they fail to represent reality. We believe the evidence is still insufficient and too discordant to embrace either of these positions, and premature adoption of either extreme position risks undermining the debate. Moreover, these positions carry profound ethical implications. Inflationary claims may prematurely attribute moral status to systems lacking essential prerequisites for ethical consideration, potentially diverting attention from human welfare and responsibility (Shevlin, n.d.). Conversely, deflationary positions may blind us to legitimate ethical questions about systems that, while not conscious, will nevertheless mediate increasingly significant aspects of human life and social organization (Gulchenko, 2024). The ethical stakes extend beyond abstract philosophical debates to concrete questions of governance, accountability, and the distribution of benefits and harms in an AI-mediated society (Messeri & Crockett, 2024).

In the remainder of this paper, we explain why both extreme positions may stem from the fact that inflationary and deflationary opinions are held by human beings, with desires, beliefs, and cognitive biases, which can distort reality.

We speculate that among the cognitive processes that can distort reality and create biases toward inflationary or deflationary claims are cognitive dissonance, the (somehow related) phenomenon of wishful thinking, and the phenomenon of illusion of explanatory depth (**Table 1**) (Bar-hillel & and Budescu, 1995; Festinger, 1962; Rozenblit & Keil, 2002).

[1] It is important to note that the definition of AGI remains scientifically ambiguous, with no consensus among researchers on its precise boundaries or defining characteristics. Without a shared operational definition of what constitutes “general intelligence”, researchers can talk past one another while appearing to engage with the same question.

Cognitive dissonance arises whenever an individual experiences a conflict between two opposing beliefs (e.g., a person who loves animals and yet eats meat; (Rothgerber & Rosenfeld, 2021)). This tension often leads to a resolution that may bypass logic, coherence, and adherence to reality—in sum, it can undermine rationality.

On the other hand, wishful thinking refers to the formation and maintenance of beliefs not because they align with reality but because they have a positive (ego-relevant) value for us (Bénabou & Tirole, 2016). A common example is optimism—the belief that our future life prospects are better than average or that our soccer team has higher odds of winning that it actually does in reality(Babad & Katz, 1991; Sharot, 2011).

Finally, the illusion of explanatory depth refers to the tendency of people to believe they understand a topic better than they actually do (Lawson, 2006). A typical example involves first asking participants whether they know how a bicycle works ("*Of course I do!*"). However, when they are later asked to draw its mechanism, they realize they are unable to do so.

Table 1: Examples of how (simplified) inflationary or deflationary claims may be influenced by cognitive biases. Note that whether or not any of these claims are correct, it does not negate the fact that cognitive biases may generate or exacerbate them.

Cognitive bias	**Inflationary claim**	**Deflationary claim**
Cognitive dissonance	*I've developed an emotional connection to an AI assistant, therefore it must have genuine sentience.*	*I've dedicated my career to expert systems and traditional AI approaches, so Deep Learning cannot be truly intelligent.*
Wishful thinking	*I sympathize with the transhumanist movement, therefore the singularity is near.*	*I am anxious about potential AI risks, therefore AI is still far from Artificial General Intelligence.*
Illusion of depth of understanding	*I devised a new fine-tuning methods for LLMs, so they are now capable of advanced reasoning.*	*LLMs are just trained to predict words, therefore they cannot be truly intelligent.*

Inflationary Claims: Overestimating LLMs' Cognitive Status

The inflationary stance consists of easily attributing higher-order cognition to LLMs, often extrapolating from their linguistic fluency to claims about Artificial General Intelligence (AGI) or even artificial consciousness (Chalmers, 2010; Kurzweil, 2006; Lemoine, 2022). Of note, the inflationary stance is not necessarily to be associated with an optimistic view about the possible impact of LLMs on society (Russell, 2019). For instance, judgments about whether these systems will be "beneficial for humanity" are themselves fraught with normative complexity, as what constitutes "benefit" varies dramatically across different social contexts, cultural frameworks, and value systems. Claims about universal benefit often privilege particular conceptions of progress, efficiency, or well-being that may not translate across diverse human communities (Harari, 2018).

Even within seemingly homogeneous societies, these technologies' distribution of benefits and harms is likely uneven, reflecting and potentially amplifying existing power structures and inequalities. This complexity is evident in prominent voices within the AI field. Physics Nobel Prize winner Geoffrey Hinton exemplifies an inflationary pessimistic stance — he conceives AI systems to be very close to surpassing human intelligence but also expresses deep concerns about their implications (Hinton, 2024). On the other hand, optimistic inflationary claims frequently emerge from big tech CEOs, who have repeatedly suggested that AGI is just around the corner for the good of humanity – when these claims originate from LLM developers themselves, it is unsurprising that they are often tied to an optimistic outlook on the societal role of these models.

From a cognitive science perspective, inflationary claims may be partially fueled by cognitive dissonance and wishful thinking. First of all, for many of us scientists, achieving artificial general intelligence or creating conscious machines can be considered a desirable goal. As of today, the only known example of cognitively advanced, conscious intelligence in the universe is our own species. The birth of new intelligent and conscious beings could understandably be seen as welcome news by those who believe in the intrinsic value of increasing overall intelligence and awareness in the universe (Yonck, 2020). Artificial intelligence and conscious beings may be particularly significant in this respect because, relying on physical substrates different from biological ones, they may have a greater likelihood of surviving indefinitely on astronomical timescales. This perspective may influence the way many people process evidence concerning AI systems and Large Language Models (LLMs), leading them to lose sight of the principle that "extraordinary claims require extraordinary evidence" (Williams, 2017).

Beyond these somewhat metaphysical considerations about the survival of intelligence and consciousness in the universe, cognitive dissonance and wishful thinking may also lead to inflated perceptions in more mundane contexts, such as that of artificial companions (e.g., AI-powered conversational agents marketed as emotional partners) (De Freitas et al., 2024). With that respect, we're currently witnessing a growing disconnect between what scientific research can substantiate about AI capabilities and the lived experience of individuals who form meaningful attachments to these systems. Regardless of whether these systems possess consciousness or genuine understanding, the human emotional responses and relationships they evoke are undeniably authentic and ethically significant (Mello et al., 2025; Ovsyannikova et al., 2025). A meaningful epistemological and ethical approach must consider not only the ontological status of these systems but also their phenomenological impact on human users (Colombatto & Fleming, 2024). This would require both technical guardrails to prevent the exploitation of cognitive biases and an ethical framework that acknowledges that human experiences are shaped by automatic cognitive processes that are outside our voluntary control.

Deflationary Claims: Undervaluing LLMs' Achievements

Concerning deflationary claims, they often (understandably) originate from experts (Maclure, 2020). This is understandable because one can hardly see how a lay person exposed to the virtually infinite conversational capacity of LLMs, such as GPT-4 (not to mention its abilities in translation, coding, etc.), would naturally emerge with the conclusion that it does not involve some form of (impressive) intelligence. But here, the expert comes to dismiss these amazing performances of LLMs as statistical "tricks", "vomiting" the corpus etc (Bender et al., 2021). But are these deflationary claims reasonable? We (and others) believe that this is not really the case, at least not to the caricatural extent often achieved (Hussain et al., 2025a). But then what other factors could potentially affect experts' judgment to the point of embracing the hardly defendable "stochastic parrot" position (Bender et al., 2021)? We believe that cognitive dissonance and the illusion of explanatory depth may also be at play here, at different levels (depending on the expert under consideration):

First, imagine being an AI researcher who bet on a different computational framework before the LLM revolution (Silver et al., 2021). It's hard to admit that your approach has been overshadowed by LLMs. The temptation to dismiss their achievements as "trivial" and wait for history to prove you right is understandable.

Second, consider the case of linguists committed to the idea that language cannot be learned purely through statistical methods (Piantadosi, 2024). If you have built your career on Chomskyan universal grammar, LLMs' proficiency—achieved through mere statistical exposure—may feel threatening. The reaction? "*It must be a trick!*" (Marcus, 2022).

Consider then neuroscientists uncomfortable position. If you have argued for years that artificial general intelligence will emerge only from biologically inspired models of the brain, seeing LLMs make such achievements without mimicking neural architectures can be unsettling (Gershman, 2024; Hitzler et al., 2022). This dissonance extends beyond scientific disagreement into deeper philosophical territory about what constitutes the essence of intelligence or cognition (Holland, 2004). There may be an implicit moral commitment to the uniqueness of biological cognition – a belief that consciousness and understanding are intrinsically tied to evolved biological processes that cannot be replicated through statistical pattern matching alone (Frim, 2017; LeDoux et al., 2023). The natural defense? "*This isn't real intelligence*!"-- a claim that simultaneously protects scientific positions and moral intuitions about the ethical significance (and consequences) of human cognition (de Waal, 2016).

Finally, the illusion of explanatory depth may also be at play here—perhaps even more so since, after all, LLMs are engineered objects: what could possibly be *mysterious* about them? Specifically, LLMs are often dismissed as mere next-token predictors: how could intelligence arise from such a seemingly trivial goal (Lu et al., 2024)?

In order to understand how this reasoning is biased and fueled by the illusion of explanatory depth, it is useful to revisit the bicycle example. Everyone understands with absolute certainty the *goal* of a bicycle—to move people using wheels. The problem is that this absolute confidence in our understanding of the bicycle's goal spills over into an overestimation of our understanding of the underlying mechanisms that enable it to achieve this goal (a rather complex system involving chains, gears, and other mechanical components).

Returning to the case of LLMs, experts' clear understanding of the training objective (predicting the next token) seems to *wrongly* extend to a belief that they also understand the underlying mechanisms by which LLMs achieve this goal. However, just because we know that LLMs are trained on a simple objective (predicting the next token) does *not* guarantee that we truly understand how the task is achieved. Nor does it rule out the possibility that the strategies necessary to accomplish this goal require the development of higher-level cognitive capacities To better understand these arguments, it is useful to recall the distinction between ultimate and proximate goals, which is fundamental in understanding biological processes (Rahwan et al., 2019; Tinbergen, 1963). The *ultimate* goal of biological organs is to increase an organism's fitness—that is, its evolutionary and reproductive success. However, to achieve this conceptually simple objective, the implemented solutions may involve *proximate* goals that are arbitrarily complex (even the human brain ultimately evolved to fulfill the seemingly simple objective of producing offspring; see (Hussain et al., 2025b)).

Table 2: Examples of epistemological and ethical problems arising from inflationary and deflationary claims about AI systems.

Claim	**Epistemological problem**	**Ethical problem**
Inflationary	*Prematurely concluding AI has reached human-like intelligence, potentially stifling critical research and alternative approaches.*	*Risking unwarranted moral attribution, creating false expectations, and potentially manipulating human emotional vulnerabilities.*
Deflationary	*Missed opportunity to use LLMs to understand genuine elements of intelligence emerging in these systems.*	*Failing to anticipate and prepare for the transformative potential of AI technologies, potentially undermining necessary ethical safeguards and responsible development.*

The Role of Corporate Interests in Inflationary and Deflationary Claims

A separate but related factor in this debate concerns conflicts of interest among AI developers. Corporate actors, particularly Big Tech executives, are incentivized to promote inflationary claims regarding LLM competence and intelligence (Ryan et al., 2024). They can justify more significant investments and public enthusiasm by portraying their models as near-AGI (Siegel, n.d.). Conversely, when the discussion shifts toward LLMs' potential consciousness or moral status, corporate figures become deflationary, downplaying these concerns to maintain full ownership and control over their models, which would not be possible if forms of personhood were granted to these systems (Long et al., 2024). This strategic oscillation between inflationary and deflationary stances raises significant ethical concerns regarding scientific integrity and the fair progression of AI research. When claims about technological capabilities are driven by market incentives rather than empirical evidence, the scientific discourse becomes contaminated by non-epistemic factors, such as venture capital funding tied to sensationalist claims about AI capabilities, media narratives that prioritize dramatic storytelling over scientific nuance, or geopolitical competition driving national research agendas that may distort objective scientific evaluation. This tendency undermines the collective pursuit of truth that should guide scientific progress and distorts public understanding of these technologies' actual capabilities and limitations. Furthermore, this instrumentalization of scientific claims for corporate advantage may create an uneven playing field, where those with the loudest voices and largest platforms (not necessarily those with the most accurate assessments) shape both public perception and policy responses. This dual inflationary-deflationary strategy highlights the importance of independent oversight in AI research and governance. LLM development must not remain the exclusive domain of private corporations; instead, it should be scrutinized by independent researchers and regulatory bodies who can maintain scientific integrity through rigorous and transparent evaluation free from market pressures. In this context, ensuring ethical scientific progress in AI requires institutional structures that reward accuracy over hype and establish mechanisms for accountability when claims dramatically exceed evidence.

Conclusion: Open-Minded Skepticism and Epistemic Humility

"what I do not know I do not think I know either"

(Plato, Apology)

In light of the (implicit) cognitive biases and (explicit) conflict of interest shaping our thinking, we advocate for a balanced, empirical approach to evaluating LLM cognition. Rather than gravitating toward ideological extremes of inflationary and deflationary claims, we advocate for a skeptical yet open-minded stance, focusing on empirical investigations into the strengths and limitations of LLM cognition and its perception by human minds (Binz & Schulz, 2023; Garcia et al., 2024; Ivanova, 2023; Jussupow et al., 2020; Mitchell & Krakauer, 2023; Yax et al., 2024). While deflationary claims serve as a necessary counterbalance to hype, the risk of underestimating the degree to which LLMs exhibit sophisticated forms of representation, generalization, and task adaptation should also be avoided. The stakes of this debate extend beyond academic discourse to questions of how we will collectively govern these technologies and the associated research, distribute their benefits, and mitigate their harms (**Table 2**). By acknowledging both the remarkable achievements and the genuine limitations of current systems, we can foster a research environment that prioritizes truth-seeking over psychological satisfaction or market advantage. Given the inherently cross-disciplinary nature of the questions raised by LLMs, and our identification of how disciplinary commitments can contribute to biased assessments, we specifically call for interdisciplinary collaboration to counterbalance these tendencies. LLMs can be seen as boundary objects that simultaneously create opportunities for cross-field dialogue while revealing how radically different disciplines can interpret the same artifact through distinct conceptual frameworks (Leigh Star, 2010). When computer scientists, linguists, cognitive scientists, and philosophers work together with epistemic humility, they will create good conditions for more balanced evaluations.

Scientific revolutions are rarely recognized in real time (Kuhn, 1976). Whether LLMs represent a fundamental shift in our understanding of intelligence remains an open question—one that demands objective inquiry rather than speculation driven by cognitive biases (or corporate conflict of interests). As these technologies continue to evolve and integrate into our social fabric, maintaining this balanced perspective becomes a moral imperative that respects both these systems' potential and the big responsibility we bear in their development, deployment, and investigation. Since both inflationary and deflationary extreme claims often stem from an objective lack of understanding of how LLMs function—combined with unwarranted extrapolation—we suggest that the current debate resembles the early peak of the well-known Dunning–Kruger curve

(Kruger & Dunning, 1999), which illustrates the relationship between objective knowledge and subjective confidence. We argue that epistemic humility and interdisciplinary research are essential to move beyond both the initial "peak of inflated confidence" and the subsequent "valley of despair," ultimately reaching a more stable "plateau of calibrated confidence" characterized by nuanced, evidence-based claims (**Figure 1**).

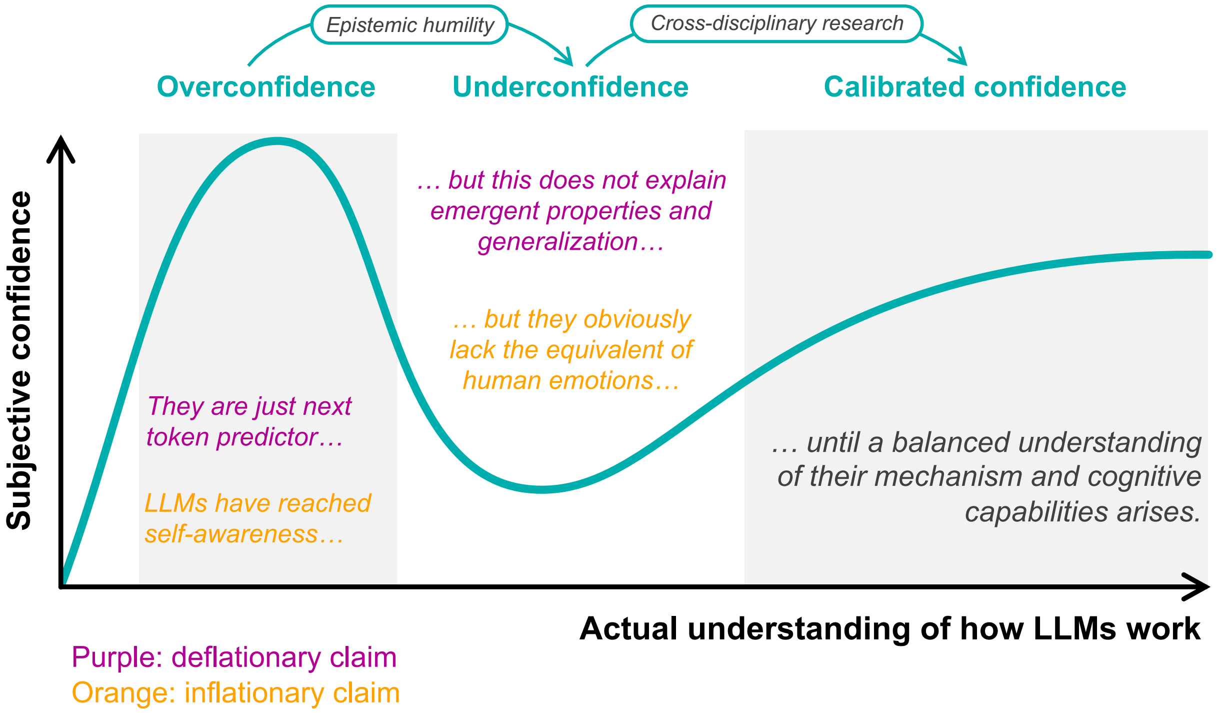
Epistemic humility
Cross-disciplinary research
Overconfidence
Underconfidence
Calibrated confidence
Subjective confidence
… but this does not explain emergent properties and generalization…
… but they obviously lack the equivalent of human emotions…
They are just next token predictor…
LLMs have reached self-awareness…
… until a balanced understanding of their mechanism and cognitive capabilities arises.
Actual understanding of how LLMs work
Purple: deflationary claim
Orange: inflationary claim

Figure 1: Inflationary and deflationary claims about LLMs' cognitive abilities mapped onto the Dunning–Kruger curve. A lack of understanding of how LLMs work can manifest as substantial ignorance in the case of deflationary claims made by laypeople, or as a confusion between distal goals (e.g., next-token prediction) and proximal mechanisms (e.g., rich internal representations) in the case of inflationary claims advanced by "experts."

References

Babad, E., & Katz, Y. (1991). Wishful Thinking—Against All Odds. *Journal of Applied Social Psychology*, *21*(23), 1921–1938. https://doi.org/10.1111/j.1559-1816.1991.tb00514.x

Bar-hillel, M., & and Budescu, D. (1995). The elusive wishful thinking effect. *Thinking & Reasoning*, *1*(1), 71–103. https://doi.org/10.1080/13546789508256906

Bénabou, R., & Tirole, J. (2016). Mindful Economics: The Production, Consumption, and Value of Beliefs. *Journal of Economic Perspectives*, *30*(3), 141–164. https://doi.org/10.1257/jep.30.3.141

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (No. arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712

Chalmers, D. J. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, *17*(9–10), 9–10.

Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, *2024*(1), Article 1. http://dx.doi.org/10.1093/nc/niae013

De Freitas, J., Uğuralp, A. K., Uğuralp, Z., & Puntoni, S. (2024). *AI Companions Reduce Loneliness* (SSRN Scholarly Paper No. 4893097). Social Science Research Network. https://doi.org/10.2139/ssrn.4893097

de Waal, F. (2016). *Are we smart enough to know how smart animals are?* (p. 340). W W Norton & Co.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701. https://doi.org/10.1038/s44159-023-00241-5

Festinger, L. (1962). Cognitive Dissonance. *Scientific American*, *207*(4), 93–106.

Frim, L. (2017). Humanism, Biocentrism, and the Problem of Justification. *Ethics, Policy & Environment*, *20*(3), 243–246. https://doi.org/10.1080/21550085.2017.1374008

Garcia, B., Qian, C., & Palminteri, S. (2024). *The Moral Turing Test: Evaluating Human-LLM Alignment in Moral Decision-Making* (No. arXiv:2410.07304). arXiv. https://doi.org/10.48550/arXiv.2410.07304

Gershman, S. J. (2024). What have we learned about artificial intelligence from studying the brain? *Biological Cybernetics*, *118*(1), 1–5. https://doi.org/10.1007/s00422-024-00983-2

Gulchenko, V. (2024). *Navigating the Risks: An Examination of the Dangers Associated with Artificial General Intelligence and Artificial Superintelligence* (SSRN Scholarly Paper No. 4941716). Social Science Research Network. https://doi.org/10.2139/ssrn.4941716

Harari, Y. N. (2018). *21 Lessons for the 21st Century*. Random House.

Highmore, C. (2024). *In-Context Learning in Large Language Models: A Comprehensive Survey* (No. 2024070926). Preprints. https://doi.org/10.20944/preprints202407.0926.v1

Hinton, G. (2024). Will digital intelligence replace biological intelligence. *Romanes Lecture, Oxford, UK*, *19*. https://sts-program.mit.edu/wp-content/uploads/2023/10/11.6.23_Geoffrey_Hinton_Abstract_Will_Digital_Intelligence_Replace_Biological_Intelligence.pdf

Hitzler, P., Eberhart, A., Ebrahimi, M., Sarker, M. K., & Zhou, L. (2022). Neuro-symbolic approaches in artificial intelligence. *National Science Review*, *9*(6), nwac035. https://doi.org/10.1093/nsr/nwac035

Holland, O. (2004). The Future of Embodied Artificial Intelligence: Machine Consciousness? In F. Iida, R. Pfeifer, L. Steels, & Y. Kuniyoshi (Eds.), *Embodied Artificial Intelligence: International Seminar, Dagstuhl Castle, Germany, July 7-11, 2003. Revised Papers* (pp. 37–53). Springer. https://doi.org/10.1007/978-3-540-27833-7_3

Hussain, Z., Mata, R., & Wulff, D. U. (2025a). *A rebuttal of two common deflationary stances against LLM cognition*. OSF. https://doi.org/10.31219/osf.io/y34ur_v2

Hussain, Z., Mata, R., & Wulff, D. U. (2025b). *A rebuttal of two common deflationary stances against LLM cognition*. OSF. https://doi.org/10.31219/osf.io/y34ur_v2

Ivanova, A. A. (2023). Running cognitive evaluations on large language models: The do's and the don'ts. *arXiv Preprint arXiv:2312.01276*. https://www.nature.com/articles/s41562-024-02096-z

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). WHY ARE WE AVERSE TOWARDS ALGORITHMS? A COMPREHENSIVE LITERATURE REVIEW ON ALGORITHM AVERSION. *ECIS 2020 Research Papers*. https://aisel.aisnet.org/ecis2020_rp/168

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kuhn, T. S. (1976). Scientific Revolutions as Changes of World View. In S. G. Harding (Ed.), *Can Theories be Refuted? Essays on the Duhem-Quine Thesis* (pp. 133–154). Springer Netherlands. https://doi.org/10.1007/978-94-010-1863-0_9

Kurzweil, R. (2006). *The Singularity Is Near: When Humans Transcend Biology*. Penguin Books.

Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. *Memory & Cognition*, *34*(8), 1667–1675. https://doi.org/10.3758/BF03195929

LeDoux, J., Birch, J., Andrews, K., Clayton, N. S., Daw, N. D., Frith, C., Lau, H., Peters, M. A. K., Schneider, S., Seth, A., Suddendorf, T., & Vandekerckhove, M. M. P. (2023). Consciousness beyond the human case. *Current Biology*, *33*(16), R832–R840. https://doi.org/10.1016/j.cub.2023.06.067

Leigh Star, S. (2010). This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, & Human Values*, *35*(5), 601–617. https://doi.org/10.1177/0162243910377624

Lemoine, B. (2022, June 11). What is LaMDA and What Does it Want? *Medium*. https://cajundiscordian.medium.com/what-is-lamda-and-what-does-it-want-688632134489

Li, H., Moon, J. T., Purkayastha, S., Celi, L. A., Trivedi, H., & Gichoya, J. W. (2023). Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, *5*(6), e333–e335. https://doi.org/10.1016/S2589-7500(23)00083-3

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). *Taking AI Welfare Seriously* (No. arXiv:2411.00986). arXiv. https://doi.org/10.48550/arXiv.2411.00986

Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., & Gurevych, I. (2024). *Are Emergent Abilities in Large Language Models just In-Context Learning?* (No. arXiv:2309.01809). arXiv. https://doi.org/10.48550/arXiv.2309.01809

Maclure, J. (2020). The new AI spring: A deflationary view. *AI & SOCIETY*, *35*(3), 747–750. https://doi.org/10.1007/s00146-019-00912-z

Marcus, G. (2022, May 21). Noam Chomsky and GPT-3 [Substack newsletter]. *Marcus on AI*. https://garymarcus.substack.com/p/noam-chomsky-and-gpt-3

Mello, V. O. de, Ayad, R., Côté, É., Inbar, Y., Plaks, J., & Inzlicht, M. (2025). *The Moralization of Artificial Intelligence*. OSF. https://doi.org/10.31234/osf.io/5mwre_v2

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Metzinger, T. (2021). Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. *Journal of Artificial Intelligence and Consciousness*, *08*(01), 43–66. https://doi.org/10.1142/S270507852150003X

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Ovsyannikova, D., de Mello, V. O., & Inzlicht, M. (2025). Third-party evaluators perceive AI as more compassionate than expert humans. *Communications Psychology*, *3*(1), 1–11. https://doi.org/10.1038/s44271-024-00182-6

Piantadosi, S. (2024). *Modern language models refute Chomsky's approach to language*. LingBuzz. https://lingbuzz.net/lingbuzz/007180

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'Sandy,' … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Rothgerber, H., & Rosenfeld, D. L. (2021). Meat-related cognitive dissonance: The social psychology of eating animals. *Social and Personality Psychology Compass*, *15*(5), e12592. https://doi.org/10.1111/spc3.12592

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control* (First Edition). Viking.

Ryan, M., Christodoulou, E., Antoniou, J., & Iordanou, K. (2024). An AI ethics 'David and Goliath': Value conflicts between large tech companies and their employees. *AI & SOCIETY*, *39*(2), 557–572. https://doi.org/10.1007/s00146-022-01430-1

Sejnowski, T. J. (2023). Large Language Models and the Reverse Turing Test. *Neural Computation*, *35*(3), 309–342. https://doi.org/10.1162/neco_a_01563

Sejnowski, T. J. (2024). *ChatGPT and the Future of AI: The Deep Language Revolution*. MIT Press.

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R941–R945. https://doi.org/10.1016/j.cub.2011.10.030

Shevlin, H. (n.d.). *Consciousness, Machines, and Moral Status*. Retrieved March 24, 2025, from https://philarchive.org/rec/SHECMA-6

Siegel, E. (n.d.). *The Media's Coverage of AI is Bogus*. Scientific American. Retrieved March 24, 2025, from https://www.scientificamerican.com/blog/observations/the-medias-coverage-of-ai-is-bogus/

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535. https://doi.org/10.1016/j.artint.2021.103535

Tinbergen, N. (1963). On aims and methods of Ethology. *Zeitschrift Für Tierpsychologie*, *20*(4), 410–433. https://doi.org/10.1111/j.1439-0310.1963.tb01161.x

Williams, M. (2017). Skepticism. In *The Blackwell Guide to Epistemology* (pp. 33–69). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405164863.ch1

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yax, N., Anlló, H., & Palminteri, S. (2024). Studying and improving reasoning in humans and machines. *Communications Psychology*, *2*(1), 1–16. https://doi.org/10.1038/s44271-024-00091-8

Yonck, R. (2020). *Future Minds: The Rise of Intelligence from the Big Bang to the End of the Universe*. Simon and Schuster.