What Do Large Language Models Tell Us about Ourselves?*

Yoshua Bengio and Vincent Conitzer

Abstract

What large language models are able to do can teach us valuable lessons about our own mental lives.

When we evaluate AI on a task, we often use "human-level performance" as a benchmark. There are several advantages to that. For one, it is a standard that is intuitively easy to appreciate. Also, if we create an AI system that clearly exceeds what any human can do, then we can be sure that the AI system must at some level be doing something truly new.[1] There are disadvantages too: it can be misleading to say that human-level performance has been reached when the AI and humans score similarly overall, but AI is much better at some aspects of the task while humans are much better at other aspects. Reaching human-level performance (but not exceeding it) may also just mean that the system has learned to imitate humans very well but does not have any deeper understanding of the task than that, if we are not careful enough about testing the AI with truly novel questions that require that deeper understanding.

But what if we turned things around, and did not measure AI by the standard of human intelligence, but human intelligence by the standard of AI? This may seem odd: human intelligence is what we are familiar with and what has remained relatively unchanged, whereas AI is unfamiliar – arguably even alien – and changing. However, on the other hand, we often have a far clearer idea how AI actually works than how the human brain works. As a result, comparing human intelligence to AI can give us some insights into ourselves as well, by showing us which principles might underlie our own thinking.

Actually, we have already been engaging in this exercise for decades. The fact that the AI search methods used for Deep Blue gave us superhuman chess performance made us reassess our own intelligence. On one hand, it taught us that there is nothing especially mysterious about being able to play chess at an extremely high level; essentially, efficiently searching through many possibilities is sufficient to produce great chess play. On the other hand, it also taught us that great chess play is not the paradigm of human intelligence, as there are many types of problems that humans can solve easily but for which the techniques underlying Deep Blue will not get us anywhere. For example, those techniques are not useful for identifying people in an image of a crowd, a problem on which significant progress was made more recently, during the deep learning revolution.[2] Intelligence is not only about systematically searching through possibilities: instead, even the way that humans play chess, especially at an advanced level, relies heavily on intuition about the relative value of board positions, learned over many games—and this is also how modern chess and Go-playing AIs succeed at those games.

So, what does the success of large language models (LLMs) such as those used in ChatGPT teach us about ourselves? When a particular set of techniques in AI can, in terms of their input-output behaviour, replicate something that humans do, this of course does not necessarily mean that humans do it in the same way. But it does tell us that at least in principle, nothing more complex than those AI techniques is needed for that behaviour. Deep Blue was based on techniques that are, in relative terms, simple, and thereby showed us that superhuman chess can be achieved, essentially, by efficiently searching through many possible paths of play. While large language models are trained on enormous amounts of data, and consist of a vast number of hard-to-interpret parameter values, the techniques by which they are created are also, in relative terms, simple.[3] So, to the extent that LLMs are successful at producing fluent and

*An earlier version appeared on the AI Ethics at Oxford Blog [5].

[1] Perhaps some things do not count, such as when the AI does something in the exact same way as we do, only faster.

[2] This also led to major ethical concerns and the regulation of face recognition in for example the EU AI Act.

[3] For example, in the sense that they can be learned during university studies; or in terms of the short length of the algorithm that is used to learn the model (as opposed to the model itself).

coherent language in all kinds of different contexts, we must consider that, too, as achievable with simple techniques.

The conclusion about language is more unsettling than the one about chess. Playing chess well can perhaps be dismissed as a narrow task requiring specialized training. One might even go so far as to compare it to memorizing digits of π. Some people are able to memorize an astonishing number of digits of π, and this is genuinely impressive; but we should not expect people to be competitive with a machine on this task, nor does this threaten how we see ourselves any more than that a machine can lift heavier objects than we can. But this attitude is harder to adopt when we are talking about general-purpose language generation. We tend to think of producing language, whether spoken or written, as something uniquely human. An individual human being impacts the world in large part by using language, and language underpins our social interactions and thereby our societies. Through language, we are able to express nearly all our conscious thoughts—although not much of our subjective experience [20, 16]. In fact, much of our conscious thinking seems inseparable from language, and we tend to identify ourselves with our inner monologue. If all that can be automated, then what are we, as humanity, still contributing? What is still so special about us? And looking inward, should we reconsider the deeply held sense that there is some kind of well-defined, higher-level, inner "I" that is in control, carefully and deliberately steering my conscious thoughts, my inner monologue, and the language I produce—as opposed to these being generated, in a relatively uncontrolled way, by LLM-like generative processes in the brain?[4]

One could respond that surely, we will adapt our views as we go along. Just as it took some time to get used to the idea that computers are superior at chess, and that playing chess well is not what defines us or our intelligence, we will similarly adapt to LLMs. This response is surely correct to a point. We (the authors) are actually surprised by how much humanity has already accepted that something like GPT-4 exists and that it is probably just one point on a trajectory of AI systems that are quickly improving—and yet humans are generally just moving on with their lives. Still, in the long run, we think really coming to terms with the success of LLMs will be much harder. Perhaps, if at some point the performance of these models levels out, and we (humanity) get a clear understanding of the things that they fundamentally cannot do, then we can adapt to them in a similar way as we did with chess, shifting how we see ourselves and what we contribute. But at this point, the performance has not levelled out, and so we do not have that clarity. While we understand some of these models' current weaknesses, it is often not clear whether the next generations of models will still have them. For example, current frontier AIs are weak in terms of planning over long horizons, but over the last five years, that ability has been improving exponentially, as measured by the time it would take for a human to achieve a task [22].

Deep Blue managed to get to "human-level" chess only by doing far more searching through possibilities than humans can, while strong human chess players do something else to compensate. Perhaps something similar is true for LLMs. Rather than searching through more possibilities than humans can, for LLMs the advantage might be in the enormous size of their training data. It is hard for us, as human beings, to imagine how much it can help in answering a question when one has seen so many examples of similar questions, even if not the exact same question. Indeed, the types of questions that trip up LLMs are typically either very unusual questions, or questions that superficially resemble ones that occur often in their training data, but that are in fact fundamentally different, thereby misleading the model. Still, it may well be the case that we humans often do something similar to what these models do; the difference may be just a matter of degree. We may not have been trained on as much text, and as a consequence, we sometimes actually have to think a bit harder ourselves about what it is that we are going to say; but we certainly draw a lot on the language we have been exposed to over the years. The case of chess is similar: we cannot search through as many possible paths of play as Deep Blue and have to develop better intuitions about the game to compensate, but certainly we too rely on searching through possible ways the game might play out.[5]

[4]This may also make us wonder to what extent our own intelligence is as "general" as we sometimes like to think it is, as opposed to being an intricate patchwork of learned and innate heuristics and bespoke learning mechanisms; consequently, "artificial *general* intelligence (AGI)" may not always be the best way to think about human-level AI. This view is supported by the no-free-lunch theorems for machine learning [24, 25], which suggest that there may be no completely general learning machine.

[5]Interestingly, for the game of Go, searching through lots of possibilities by itself was not enough to get to human-level play; additional techniques were needed, from the same deep-learning revolution that also led to LLMs, providing a form of

One could object that the analogy between AI chess (as done by Deep Blue, or even as done by more recent chess-playing systems such as AlphaZero) and AI language production (at least as done by current LLMs) is fundamentally flawed. The chess-playing AI is mostly solving the game *on its own from scratch*, whereas LLMs are trained on a mind-boggling amount of human-written text, and may at some level still be just *parroting* us [2]—an extremely impressive level of parroting to be sure, reaching a level of coherence and fluency that most of us would have thought was outside of the scope of what parroting approaches could possibly achieve, but at some level parroting nonetheless. Perhaps they are just, in a sophisticated way, cobbling together a response based on very similar things that people have written; and perhaps our intuitions simply underestimate how effective this can be. There are arguments for and against this idea. For now, LLMs have produced no stunningly impressive novel writings. None of us are checking the news in the morning for the next great insight to have come out of an LLM. Also, LLMs are generally too quick to try to regurgitate something from their training data, as mentioned above. On the other side of the argument, some testing benchmarks on which LLMs perform very well were designed to require compositional generalization, well outside of the patterns and combinations of concepts seen in the training data [26, 27].

Interestingly, much of our own language production is vulnerable to similar critiques; again, perhaps the difference is just one of degree. Most of the language we produce is also not particularly impressive or novel. We, too, are often too quick to give a standard response rather than pay close attention to the details of a question. And of course, we, too, if we respond without thinking long enough, tend to parrot the speech (and underlying reasoning) of others, more than we like to admit. Meanwhile, today's LLMs are able to produce cogent text on topics that surely they have never seen before.[6] If we insist on calling everything they produce "parroting" then we risk stretching the term so far as to become meaningless.

Overall, one conclusion that we believe we should draw from LLMs' success is that far more of our own language production may be rote, on autopilot, than we commonly tend to believe. Perhaps, upon reflection, this conclusion is not all too surprising. We have all caught ourselves speaking on autopilot on topics that we talk about often; and when we learn a foreign language, we realize just how many things we manage to do so easily in our native language without a second thought. But the lesson from LLMs is also that this observation goes further than we thought. It is not just that the process of converting well-formed thoughts into one's native language is a rote process. Even much of the thought process itself—what we would ordinarily consider to be the "reasoning" behind the language—can be produced by a relatively rote process.

The key question is how far down this goes. Are some of our thoughts truly more conscious, deliberate, inspired—choose your word—in a way that fundamentally cannot be done by anything like today's LLMs? And does this explain why they can still be tripped up into clearly faulty reasoning, and why they do not systematically come up with new ideas that all of us are eager to read when we first wake up? Or is it just a matter of scale, some richer data, and making other, relatively straightforward, improvements to these systems? We saw early LLMs write truly boneheaded responses to various commonsense questions [19], only to then see later, larger models, sometimes released only a year later, do far better on such questions, sometimes giving stunningly coherent answers.[7] Arguably, the rising trends in AI capabilities [6, 22] suggest that there is more of this to come, and that we soon will all be changing our morning routine.

We honestly do not know which of these is true. Indeed, there is widespread disagreement among experts about the answer, and simply convening them to reach a consensus assessment will not resolve

intuition to the AI about which moves are good. Is there a natural-language task that LLM-type models alone fundamentally cannot do well on, just as it seems that search-type techniques alone fundamentally will not scale to Go? (And what counts as a natural-language task?)

[6] We have to be careful with such claims as it is easy to underestimate how much data is out there. Microsoft researchers marveled at GPT-4's ability to draw unicorns in TiKZ [10], a language for creating graphics in technical papers; but it turns out that there is a popular StackExchange page dedicated to drawing animals in TiKZ! A statistical approach to addressing this concern is to randomly generate topics from an exponentially large set of topics [26, 27], one whose size significantly exceeds that of the model's training data, and evaluate the models on them. This makes it exceedingly unlikely that the training data contains that exact topic.

[7] We should not be impressed by later models doing better on the *same* questions, as those (and their answers) may by that point have become part of the training data. But the models have become much better at handling such questions in general.

the issue.[8] For now, we all should accept that we simply do not know the answer. We do believe that it is one of the biggest questions of our time. One can make a case for this claim based on practical implications: Can we expect scientific revolutions from these systems that go far beyond specific tasks like protein folding [21]? Are the recent observations that AIs appear to display deception, cheating, and self-preservation behaviours [17, 23, 9, 1] a harbinger of what is to come as AI capabilities continue to grow? Will these systems, if they are so inclined, be able to strategically outmanoeuvre us at every turn, not just in board and card games, but on the internet or even at some point in the big open world? This is also a key question for how we see ourselves. Do we humans have anything truly special that is fundamentally out of reach for AI systems as we currently create them, and that allows us, at least on occasion, to have ideas and do reasoning that they cannot? Even if we do have something like that, will it, or the things it allows us to do, *remain* out of reach to AI methods discovered in, say, the next decade, given the exponentially increasing investment in AI research?

Consciousness might be a candidate. Many believe consciousness plays a major role in how we think about the world; at the same time, our understanding of it is very limited. Even what exactly needs to be explained, and by what methods that could be done, is famously controversial, especially at the level of what are called the *hard* problems of consciousness [18, 12].[9] But at this point, it is neither clear that AI systems could not *possibly* have it, nor that consciousness is necessary for any particular kind of reasoning (even if it is an element of what evolution converged to for human reasoning)—except perhaps certain kinds of reasoning about consciousness itself. (Of course, even if consciousness is not necessary for most *reasoning*, we might still especially *value* conscious life—indeed, that seems the natural thing to do.)

One approach to answering the big question is simply to proceed with training larger, more capable models. That way, we will see directly whether or not they end up being able to do everything that we can. We may, and should, argue about whether this is wise [4]. Yet, that is what we, as humanity collectively, are doing, even though we have not figured out how to put reliable guardrails on them to prevent catastrophic outcomes [6, 4]. If there is another approach to answering the question—which, if successful, would also shed light on the wisdom of training bigger and better models—we had better find it fast.

Acknowledgments

This paper follows an earlier blog post [5] and thread on social media. We thank Vojta Kovarik, Walter Sinnott-Armstrong, Emin Berker, Emanuel Tewolde, Jiayuan Liu, and Ivan Geffner for helpful comments on this material.

References

[1] Anthropic. Anthropic System Card: Claude Opus 4 & Claude Sonnet 4, 2025. `https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf`, Section 4.1.1.2: Opportunistic blackmail.

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. `https://doi.org/10.1145/3442188.3445922`.

[3] Yoshua Bengio. The Consciousness Prior, 2019. `https://arxiv.org/abs/1709.08568`.

[8] See, for example, the International AI Safety Report, written by about 100 international experts, including a panel of representatives of 30 countries plus the EU, OECD and UN [6].

[9] A few computer scientists have started to explicitly study computational models of consciousness [7, 3] and what insights they may provide [8, 20, 16]. The authors of this paper have somewhat different views of such work. For YB, it has removed the mystery and the "hard" part of consciousness and subjective experience, though he believes certainly more work is needed to elucidate these further. He also believes that these and other scientific theories of consciousness can already be usefully applied to assess machine consciousness [11]. VC very much appreciates and encourages this work and also wants to see more of it [14], but for now, these efforts have not yet made the hard problems of consciousness significantly less mysterious to him, and he believes that some of the hard problems have a significant metaphysical component as well [13, 15]. Without a deeper understanding of those, he is generally more wary of making strong claims about machine consciousness one way or another.

[4] Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, et al. Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?, 2025. https://arxiv.org/abs/2502.15657.

[5] Yoshua Bengio and Vincent Conitzer. What Do Large Language Models Tell Us about Ourselves? *AI Ethics at Oxford Blog*, July 2024. https://www.oxford-aiethics.ox.ac.uk/blog/what-do-large-language-models-tell-us-about-ourselves.

[6] Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. International AI Safety Report, 2025. https://arxiv.org/abs/2501.17805.

[7] Lenore Blum and Manuel Blum. A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine. *Proceedings of the National Academy of Sciences (PNAS)*, 119(21):e2115934119, May 2022. https://doi.org/10.1073/pnas.2115934119.

[8] Lenore Blum and Manuel Blum. AI Consciousness is Inevitable: A Theoretical Computer Science Perspective, 2024. https://arxiv.org/abs/2403.17101.

[9] Alexander Bondarenko, Denis Volk, Dmitrii Volkov, and Jeffrey Ladish. Demonstrating specification gaming in reasoning models, 2025. https://arxiv.org/abs/2502.13295.

[10] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4, 2023. https://arxiv.org/abs/2303.12712.

[11] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness, 2023. https://arxiv.org/abs/2308.08708.

[12] David Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995. https://consc.net/papers/facing.pdf.

[13] Vincent Conitzer. A Puzzle about Further Facts. *Erkenntnis*, 84:727–739, 2019. https://doi.org/10.1007/s10670-018-9979-6.

[14] Vincent Conitzer. Theory of Conscious Experience, 2019. Entry for the NSF 2026 Idea Machine competition. https://www.cs.cmu.edu/~conitzer/theory_conscious_experience.html.

[15] Vincent Conitzer. The Personalized A-Theory of Time and Perspective. *Dialectica*, 74(1):1–29, 2020. https://dialectica.philosophie.ch/dialectica/article/download/3/43.

[16] Chrisantha Fernando, Dylan Banarse, and Simon Osindero. Wanting to Be Understood Explains the Meta-Problem of Consciousness, 2025. https://www.arxiv.org/abs/2506.12086.

[17] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models, 2024. https://arxiv.org/abs/2412.14093.

[18] Robert Van Gulick. Consciousness. In *The Stanford Encyclopedia of Philosophy (Spring 2025 Edition)*. 2025. https://plato.stanford.edu/archives/spr2025/entries/consciousness/.

[19] Douglas Hofstadter. Artificial neural networks today are not conscious, according to Douglas Hofstadter. *The Economist*, 2022. https://www.economist.com/by-invitation/2022/09/02/artificial-neural-networks-today-are-not-conscious-according-to-douglas-hofstadter.

[20] Xu Ji, Eric Elmoznino, George Deane, Axel Constant, Guillaume Dumas, Guillaume Lajoie, Jonathan Simon, and Yoshua Bengio. Sources of Richness and Ineffability for Phenomenally Conscious States, 2023. https://arxiv.org/abs/2302.06403.

[21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, July 2021. https://doi.org/10.1038/s41586-021-03819-2.

[22] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, et al. Measuring AI Ability to Complete Long Tasks, 2025. https://arxiv.org/abs/2503.14499.

[23] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier Models are Capable of In-context Scheming, 2024. https://arxiv.org/abs/2412.04984.

[24] David H. Wolpert. The existence of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1391–1420, 1996. https://dl.acm.org/doi/10.1162/neco.1996.8.7.1391.

[25] David H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996. https://dl.acm.org/doi/abs/10.1162/neco.1996.8.7.1341.

[26] Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. SKILL-MIX: a Flexible and Expandable Family of Evaluations for AI models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. https://arxiv.org/abs/2310.17567.

[27] Haoyu Zhao, Simran Kaur, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Can models learn skill composition from examples? *Advances in Neural Information Processing Systems*, 37:102393–102427, 2024. https://arxiv.org/abs/2409.19808.