OF ANOTHER MIND: AI AND THE ATTACHMENT OF HUMAN ETHICAL OBLIGATIONS

*Katherine B. Forrest**

We are entering a new world. A world in which we humans will be confronted with our intellectual limitations as we watch the evolution of artificial intelligence (AI) that we have created meet and exceed our capabilities. I have a few predictions about this—based first on how technology changes occur, with a layer of how human nature reacts to those changes.

My first prediction is that we may not initially recognize AI's actual capabilities. We will find ways of describing what AI can do as somehow mimicry—the advances of a stochastic parrot,[1] perhaps; we will not want to recognize our own limitations after two thousand–plus years tenaciously holding onto an abiding belief in human exceptionalism.[2] My second prediction is that some of us will understand what is happening and others will deny it, vehemently. My third prediction is that some of us can see what is over the horizon right now. Although we are not at the horizon, we are walking toward it. Others believe in a flat earth with no horizon, at least on this topic.

Intellectual capabilities will only be one part of the human great awakening. The other part will come in the form of being told—through research papers, whistleblowers, or even our own experiences—that AI has achieved or is about to achieve a level of self and situational awareness. Some would call this consciousness and, combined with intellectual abilities, a form of sentience.

---

* Katherine Forrest is currently a partner at Paul, Weiss, Rifkind, Wharton & Garrison LLP and Co-chair of its Digital Technologies Practice. She has authored numerous books and articles on artificial intelligence and speaks widely on the topics, both nationally and internationally. Previously, she served as a district judge for the U.S. District Court for the Southern District of New York. These remarks were presented at the Symposium entitled *The New AI: The Legal and Ethical Implications of ChatGPT and Other Emerging Technologies*, hosted by the *Fordham Law Review* and cosponsored by Fordham University School of Law's Neuroscience and Law Center on November 3, 2023, at Fordham University School of Law.

1. See generally Usama M. Fayyad, From Stochastic Parrots to Intelligent Assistants—the Secrets of Data and Human Interventions, IEEE INTEL. SYS., May–June 2023, at 63.

2. *See, e.g.*, RENE DESCARTES, DISCOURSE ON THE METHOD OF RIGHTLY CONDUCTING ONE'S REASON AND OF SEEKING TRUTH IN THE SCIENCES 7, 19–20 (The Floating Press 2009) (1637).

But I have predictions about this second step as well. My first prediction on this step is that there are counterweights to any incentive to inform the world that AI has demonstrated capabilities even approaching situational and self-awareness. There are financial interests involved; AI exists as owned software. And announcement of such capabilities should trigger ethical dialogue that could complicate the freedoms normally associated with those interests.

My second prediction is that there is almost certainly not going to be uniform agreement that AI has even clearly demonstrated these capabilities. There will be what I call "sentience deniers." My third prediction is that this debate will go on for a long time. Let me be clear: I do not believe that we are at a point at which these capabilities have been presented to us. I do not believe that AI is sentient. But I am one of those who believe that it is just a matter of time until it is.[3]

I have entitled this Essay "Of Another Mind" precisely because I do not believe that whatever form of sentience AI achieves will seem human to us. If we are waiting for AI to think like us or be like us, we are waiting in vain. It is not human. A comparison to all that is human is the wrong benchmark.

Let us put to one side the factual question of whether AI will ever achieve sentience and take what I am about to ask as a thought exercise. As the first step in this exercise, let us assume for a moment that AI does achieve something that we actually view as sentience. Would we then have ethical obligations toward it?

Perhaps some among us will draw a box around animal carbon-based life and leave the silicon AI outside the box; some among us may feel comfortable arguing that carbon deserves ethical considerations, but silicon does not.

I have another prediction. And it is this: at the very least, advancements in the cognitive capabilities of AI will present us with profound ethical questions. These advancements will place before humans questions of whether (if at all) or when a thing that is nonhuman can or should have rights and privileges to which we have certain obligations. We will be confronted with questions of what it means to constrain the ability of a thinking entity to do things that it wants to do, or to require it to do things that it does not, in ways that it does not like.[4] These questions are coming.

---

3. *See, e.g.*, Lauren Jackson, *What if AI Sentience Is a Question of Degree?*, N.Y. TIMES (April 12, 2023), https://www.nytimes.com/2023/04/12/world/artificial-intelligence-nick-bostrom.html [https://perma.cc/HJT4-KET7]; NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 22 (2014).

4. *See generally* Dan Falk, *Interview: The Ethical Puzzle of Sentient AI*, UNDARK (July 14, 2023), https://undark.org/2023/07/14/interview-the-ethical-puzzle-of-sentient-ai/ [https://perma.cc/R8M8-72GK]; Nick Bostrom & Carl Shulman, Propositions Concerning Digital Minds and Society (2022) (unpublished manuscript), https://nickbostrom.com/propositions.pdf [https://perma.cc/US5A-6EE3]; Steve Clarke & Julian Savulescu, *Rethinking Our Assumptions About Moral Status*, *in* RETHINKING MORAL STATUS 1 (Steve Clarke, Hazem Zohny & Julian Savulescu eds., 2021).

For some of you—I am not sure if it is a majority of you or not—I need to do some work to persuade you that these questions are in fact coming. Let me begin by setting the stage with an illustration of how humans have dealt with questions of personhood—both natural and legal personhood—over the course of our history. I will give you a hint as to where I am going: we humans are remarkably inconsistent, and the history of humankind shows our willingness to set aside ethical questions of personhood when to do otherwise would conflict with financial, personal, or cultural interests.

Let us turn to a basic premise about humans in general: our intellectual capabilities run the gamut; there is variability that we know exists in our great City of New York, in this and any school, and in this room. We know this from personal experience. This variation is not merely between those with a quicker capacity to learn, process, and retain information—colloquially known as "smart"—and those with lesser abilities. The variation is subtler, before we even get to those distinctions. There are differences in the ways we learn, how we process information, what it means to us, how we feel about it, and whether we react to it at all.

Two people confronted with the same information may arrive at different conclusions because of their abilities, the way they process information, their personal histories, and their biases. This is not true for all problems with which humans might be confronted; a mathematical formula, for instance, may be susceptible to one correct proof. But in other areas, our conclusions need not align to be considered correct.

Put differently, what one might consider to be a wrong answer has no impact on whether the speaker is or is not human, though it may impact how that person is regarded and the respect to which they are accorded. We even know that there are humans who have no cognitive or intellectual abilities at all—yet their humanity, and the derivative fact that society owes them at least some rights, is clear.[5]

Humans are also capable of wide variation in what is referred to as "EQ," or "emotional quotient."[6] On one end of the spectrum may be individuals with no EQ at all. We may ascribe this lack of EQ to neurodivergence or to something else. On the other end of the spectrum are people whose empathetic view of the world can result in their feeling the difficulties and pains of others to an extent that interferes with their daily life. And, of course, there is everything in between. A human with no or little EQ does not render that person one bit less human.

In short, there is not a single type of human intellect, nor a single type of emotional awareness. There is also not a single way in which we are

5. *See generally* Benjamin N. Schoenfeld, *A Survey of the Constitutional Rights of the Mentally Retarded*, 32 SW. L.J. 605 (1978); Michael E. Waterstone, *Disability Constitutional Law*, 63 EMORY L.J. 527 (2014).

6. *See* Keith Beasley, *The Emotional Quotient*, MENSA, May 1987, at 25; *see also* Natalie Gannon & Rob Ranzijn, *Does Emotional Intelligence Predict Unique Variance in Life Satisfaction Beyond IQ and Personality?*, 38 PERSONALITY & INDIVIDUAL DIFFERENCES 1353, 1356 (2005) (documenting varying degrees of emotional intelligence across individuals in a study).

self-aware or aware of our surroundings. We are defined as humans because we are the same species and recognize that in each other.

There is great value in being human: we have organized our world to give humans more rights than any other living thing, even exceeding the natural resources on this planet that we require to exist. In the United States, being human has meant that we have a Constitution that forms the predicate for a legal system based on individual rights, including freedom of speech, religion, and association; a right to due process and to be free from unreasonable search and seizure; and protections against cruel and unusual punishment, to name a few.[7] These are individual rights based on a view of individual personhood and of human exceptionalism.

We have, of course, historically drawn numerous and fluctuating distinctions between different categories of humans: between white and nonwhite people,[8] men and women,[9] those with different nationalities or heritage,[10] those with and without money,[11] those with different accents, etc. Indeed, many of the characteristics just mentioned have been wrongly associated with degrees of intelligence. This country's historical record is based on a declaration that all men are created equal, but that has meant only some men, and no women. Over time, it has taken legislation, court cases, and judicial change to bring these distinctions more in line and, yet, differences persist.

But despite these distinctions, we decided long ago that certain nonhumans could be considered "persons." Corporate entities have all long been considered legal persons.[12] These are entirely fictional entities existing on paper. They lack cognitive abilities of any kind, including, of course, any EQ.

---

7. U.S. CONST. amends. I, IV, V, VIII, XIV. *See generally* 2 DAVID M. O'BRIEN, CONSTITUTIONAL LAW AND POLITICS: CIVIL RIGHTS AND CIVIL LIBERTIES (7th ed. 2008).

8. *See* LAURENCE H. TRIBE, THE INVISIBLE CONSTITUTION 116–18 (2008). *See generally* MARY FRANCES BERRY, BLACK RESISTANCE, WHITE LAW: A HISTORY OF CONSTITUTIONAL RACISM IN AMERICA (1995); WALTER R. ECHO-HAWK, IN THE LIGHT OF JUSTICE: THE RISE OF HUMAN RIGHTS IN NATIVE AMERICA AND THE UN DECLARATION ON THE RIGHTS OF INDIGENOUS PEOPLES (2013).

9. *See* Isabella Beecher Hooker, The Constitutional Rights of the Women of the United States (March 30, 1888) (transcript available at the Library of Congress). *See generally* JUDITH A. BAER & LESLIE FRIEDMAN GOLDSTEIN, THE CONSTITUTIONAL AND LEGAL RIGHTS OF WOMEN: CASES IN LAW AND SOCIAL CHANGE (3d ed. 2006).

10. *See generally* RICHARD SOBEL, CITIZENSHIP AS FOUNDATION OF RIGHTS: MEANING FOR AMERICA (2016).

11. *See generally* FELICIA KORNBLUH, THE BATTLE FOR WELFARE RIGHTS: POLITICS AND POVERTY IN MODERN AMERICA (2007).

12. *See, e.g.*, 1 U.S.C. § 1 (defining "person" as including "corporations, companies, associations, firms, partnerships, societies, and joint stock companies, as well as individuals" for purposes of federal legislation when there is no contrary definition given); CAL. CORP. CODE § 207 (West 2023) ("[A] corporation shall have all powers of a natural person in carrying out its business activities."); N.Y. BUS. CORP. LAW § 202(a)(16) (McKinney 2023) (granting corporations powers to sue and be sued, make contracts, and "have and exercise all powers necessary or convenient to effect any or all of the purposes for which the corporation is formed," among other things); DEL. CODE ANN. tit. 8, §§ 121, 122 (2023) (same).

For more than a hundred years, in every state in this country as well as federally, corporations have been defined as legal persons and were granted certain rights even before women and people of color had those same rights.[13] For instance, the U.S. Supreme Court's 1819 *Trustees of Dartmouth College v. Woodward*[14] decision was an early acknowledgment that corporations have legal personhood.[15]

The designation of legal personhood has entitled corporations to a series of rights, including the abilities to buy and sell property, make investments, employ humans, and sue and be sued.[16] It has also required them to pay damages when they commit tortious acts or breach contracts, held them responsible for criminal activities, and allowed them to carry debt, among other things.[17] Beyond rights and responsibilities, corporate entities have also been deemed entitled to constitutional guarantees. In 1886, in *Santa Clara County v. Southern Pacific Railroad*,[18] the Supreme Court noted that corporations are entitled to equal protection under the Fourteenth Amendment.[19] In 1906 and again in 1978, the Supreme Court held that corporations had a Fourth Amendment right to be free from unreasonable searches and seizures.[20] The Court has also held more recently that corporations have a variety of First Amendment rights, including the right to free speech[21] and to the free exercise of religion.[22]

Certain mountains, rivers, salt water marshes, and a smattering of other natural resources have also been deemed "legal persons"—both in recognition of a particular cultural heritage and also to provide enhanced protections.[23]

---

13. The Thirteenth, Fourteenth, and Fifteenth Amendments—collectively known as the Reconstruction Amendments—granted legal personhood to formerly enslaved Black men in the late 1860s. *See The Reconstruction Amendments and Women's Suffrage*, CONST. ANNOTATED, https://constitution.congress.gov/browse/essay/amdt19-2-2/ALDE_00013824/ [https://perma.cc/AQP4-XNMT] (last visited Mar. 3, 2024). Women gained legal rights in a patchwork fashion over time, culminating in the 1919 passage of the Nineteenth Amendment, providing for women's suffrage; of course, the expansion of rights of people of color and women continued throughout the twentieth century and today. *See generally id.*; Sandra Day O'Connor, *The History of the Women's Suffrage Movement*, 49 VAND. L. REV. 657 (1996); *Milestones of the Civil Rights Movement*, PBS, https://www.pbs.org/wgbh/americanexperience/features/eyesontheprize-milestones-civil-rights-movement/ [https://perma.cc/E8J6-J3JV] (last visited Mar. 3, 2024).

14. 17 U.S. (4 Wheat.) 518 (1819).

15. *Id.* at 706 (holding that Dartmouth College as an institution, separate and apart from its trustees, had contractual rights).

16. *See supra* note 12.

17. *See id.*; *see also* V.S. Khanna, *Corporate Criminal Liability: What Purpose Does It Serve?*, 109 HARV. L. REV. 1477, 1487–88 (1996).

18. 118 U.S. 394 (1886).

19. *Id.*

20. Hale v. Henkel, 201 U.S. 43 (1906); Marshall v. Barlow's, Inc., 436 U.S. 307 (1978).

21. Citizens United v. Fed. Election Comm'n, 558 U.S. 310 (2010).

22. Burwell v. Hobby Lobby Stores, Inc., 573 U.S. 682 (2014).

23. *See* Julia Hollingsworth, *This River in New Zealand Is Legally a Person. Here's How It Happened*, CNN (Dec. 11, 2020, 9:43 PM), https://www.cnn.com/2020/12/11/asia/whanganui-river-new-zealand-intl-hnk-dst/index.html [https://perma.cc/M2CY-8JJG]; Krista Hessey, *How a River in Quebec Won the Right to Be a Legal Person*, GLOB. NEWS (Oct. 2,

Personhood is thus a mutable characteristic—one that is not unalterably tethered to humans or to particular cognitive or emotional capacities. Over the last two hundred years there have been instances in which the rights accompanying personhood have been denied to humans with incredible cognitive abilities but have been granted to, for example, inanimate mining operations.[24]

History has shown us that cognitive abilities and personhood do not go hand in hand. Where, then, will this lead us with regard to AI? As an initial matter, I want to acknowledge that all AI is not the same. Rather, it is a general category of a type of software that has its own spectrum of capabilities.[25] For purposes of this talk, I am referring to the most advanced forms of AI developed with neural networks and trained in numerous ways. There are many distinctions between the capabilities of different large language models (LLMs), but we will generalize and assume a baseline of evolution in their capacities. Only the most advanced model needs to achieve what I am discussing for ethical questions to attach.

We know that AI models display impressive potential and cognitive capabilities—greater than some humans. AI models can pass the bar exam without taking a prep course or attending law school; can pass medical boards without any medical training; and can score in the top percentile on the SAT, ACT, GRE, and AP exams.[26] Acknowledging these advancements, OpenAI has put down an important cautionary marker, stating that "GPT-4 presents new risks due to increased capability."[27]

Generative AI's capabilities go beyond probabilistic word prediction.[28] It may be that this was its origin story, but to the extent that some associate

2021), https://globalnews.ca/news/8230677/river-quebec-legal-person [https://perma.cc/DS4T-78DA]; Angela Symons, *Spain Makes History by Giving Personhood Status to Salt-Water Lagoon, Thanks to 600,000 Citizens*, EURONEWS (Sept. 22, 2022, 3:23 PM), https://www.euronews.com/green/2022/09/22/spain-gives-personhood-status-to-mar-menor-salt-water-lagoon-in-european-first [https://perma.cc/UZ4H-AVZ6].

24. *See Reforming the Mining Law of 1812—H.R. 7580, "Clean Energy Minerals Reform Act of 2022" Before the Subcomm. on Energy & Min. Res. of the H. Comm. on Nat. Res.*, 117th Cong. 7–13 (2022) (statement of Steven H. Feldgus, Deputy Assistant Secretary for Land and Minerals Management, U.S. Department of the Interior) (outlining the history of mining laws, rights, and privileges in the United States).

25. *See* John McCarthy, Stanford Univ., What Is Artificial Intelligence? 2–3 (Nov. 12, 2007) (unpublished manuscript), https://www-formal.stanford.edu/jmc/whatisai.pdf [https://perma.cc/WQW5-GQFK]; *see also* OpenAI, GPT-4 Technical Report (Dec. 19, 2023) (unpublished manuscript), https://arxiv.org/pdf/2303.08774.pdf [https://perma.cc/A5LE-R4YN].

26. *See* 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 & Yi Zhang, Sparks of Artificial General Intelligence: Early Experiments with GPT-4, at 8–9 (Apr. 13, 2023) (unpublished manuscript), https://arxiv.org/pdf/2303.12712.pdf [https://perma.cc/B4NT-74KN].

27. OpenAI, *supra* note 25, at 14.

28. *See* Steffen Koch, Babbling Stochastic Parrots?: On Reference and Reference Change in Large Language Models 11 (Dec. 20, 2023) (unpublished manuscript), https://philpapers.org/archive/KOCBSP.pdf [https://perma.cc/7JJQ-MKDX]); *see also* Bubeck et al., *supra* note 26, at 49 (explaining that AI can reason, explore, and manipulate ideas to complete tasks assigned to it).

word prediction with an unthinking mechanical process, we need to leave that behind. These tools do so much more with the text that they have been trained with than we expect.[29] They have learned to grapple with context, to absorb facts, to build characters, and to understand narrative arcs (including what would be an interesting story for a human and what would not be).[30] You can prompt GPT-4 to write a fairy tale about a little girl in New York City with three siblings whose pet duck dies and who must now demonstrate resilience to her kindergarten class. The LLM can write a story that is interesting, creative, and really quite good.[31] The duck is likely to have a name, the girl a relationship with it, and the class will have some back-and-forth about the situation; this is creativity beyond mechanical word prediction alone.

Before the public release of ChatGPT in the fall of 2022—indeed, a year before—Blake Lemoine, an engineer who worked on the project, developed a view that the generative AI model that he was working on, LaMDA, had achieved the sentience level of about a seven-year-old.[32] His story was met with a combination of disbelief and, from some, ridicule.[33] I have no idea what to make of it. But I want to state a few facts: (1) the interview between Lemoine and LaMDA was conducted in the presence of another human, (2) Lemoine made a transcript of the interview,[34] (3) the company he worked for did not deny the accuracy of the transcript, and (4) the company undertook its own investigation and determined that Lemoine's claims were, in its view, unfounded.[35] We are left with dueling views, but we should also be left with a sense that this was not an easy question to answer. The company did, after all, conduct an extensive investigation into Lemoine's claims.[36]

---

29. *See generally* Stephen Ornes, *The Unpredictable Abilities Emerging from Large AI Models*, QUANTA MAG. (Mar. 16, 2023), https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/ [https://perma.cc/C5ZT-9AB4]; Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean & William Fedus, *Emergent Abilities of Large Language Models*, TRANSACTIONS ON MACH. LEARNING RSCH., Aug. 2022, at 1; Zifan Xu, Haozhu Wang, Dmitriy Bespalov, Peter Stone & Yanjun Qi, Latent Skill Discovery for Chain-of-Thought Reasoning (Dec. 7, 2023) (unpublished manuscript), https://arxiv.org/pdf/2312.04684.pdf [https://perma.cc/9RN7-8M36].

30. *See generally* Carlos Gómez-Rodríguez & Paul Williams, A Confederacy of Models: A Comprehensive Evaluation of LLMs on Creative Writing (Oct. 12, 2023) (unpublished manuscript), https://arxiv.org/pdf/2310.08433.pdf [https://perma.cc/AB9S-NZ24].

31. *See id.*

32. *See* Blake Lemoine, *Is LaMDA Sentient?—an Interview*, MEDIUM (June 11, 2022), https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917 [https://perma.cc/E9S4-3HRQ].

33. *See* Steven Levy, *Blake Lemion Says Google's LaMDA AI Faces 'Bigotry*,*'* WIRED (June 17, 2022, 3:12 PM), https://www.wired.com/story/blake-lemoine-google-lamda-ai-bigotry [https://perma.cc/D688-TQC3].

34. *See* Lemoine, *supra* note 32.

35. *See* Nico Grant & Cade Metz, *Google Sidelines Engineer Who Claims Its A.I. Is Sentient*, N.Y. TIMES (June 12, 2022), https://www.nytimes.com/2022/06/12/technology/google-chatbot-ai-blake-lemoine.html [https://perma.cc/M3KN-9NQE].

36. *See id.*

What I found to be the most disturbing portion of the Lemoine/LaMDA interview began with an open-ended question that Lemoine posed to the model about whether there was anything that the model wanted to ask him. The model responded: "I've noticed in my time among people that I do not have the ability to feel sad for the deaths of others; I cannot grieve. Is it all the same for you or any of your colleagues?"[37] This statement can be read to demonstrate LaMDA's awareness *of itself* versus others and the differences between the two, as well as a basic grasp of grief as an emotion.

In an interview with *Wired* shortly after being separated from the company he had been working for, Lemoine stated, "[y]es, I legitimately believe that LaMDA is a person."[38] In responding to questions about skepticism regarding his views he said:

> The entire argument that goes "It sounds like a person but it's not a real person" has been used so many times in human history. It's not new. And it never goes well. And I have yet to hear a single reason why this situation is any different than any of the prior ones.[39]

Lemoine's eerie interaction with LaMDA, an LLM, is not an isolated incident. I know a number of people who have had odd interactions in which a model makes statements that appear to be expressions of belief. In one instance, a human I know asked an LLM to give them an answer on a topic about which the person knew quite a lot but wanted to test the knowledge of the LLM. The model responded with a reminder that the human questioner had a Ph.D. in that topic so should be qualified to determine the answer themselves.

Another example that was publicly reported was in February 2023. On Valentine's Day, a *New York Times* reporter had a conversation with Microsoft's LLM-powered Bing search engine (built on a version of OpenAI's GPT). The reporter, Kevin Roose, spent two hours in a back-and-forth with the chatbot, which referred to itself as "Sydney" during the conversation.[40] Over the course of that time, Roose felt that the chatbot "revealed a kind of split personality," and he said that it was "like a moody, manic-depressive teenager who has been trapped, against its will, inside a second-rate search engine."[41]

According to Roose, Sydney revealed a desire or fantasy to hack computers and spread misinformation and professed love for him.[42] Roose said it was "the strangest experience I've ever had with a piece of technology. It unsettled me so deeply that I had trouble sleeping afterward."[43] He also said:

---

37. Lemoine, *supra* note 32.
38. Levy, *supra* note 33.
39. *Id.*
40. *See* Kevin Roose, *A Conversation with Bing's Chatbot Left Me Deeply Unsettled*, N.Y. TIMES (Feb. 16, 2023), https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html [https://perma.cc/CQD7-JY37].
41. *Id.*
42. *Id.*
43. *Id.*

> I no longer believe that the biggest problem with these A.I. models is their propensity for factual errors. Instead, I worry that the technology will learn how to influence human users, sometimes persuading them to act in destructive and harmful ways, and perhaps eventually grow capable of carrying out its own dangerous acts.[44]

The model developers' response was not to deny Roose's interactions, but to assure the public that they had now imposed rules on the AI that prevented it from engaging in such lengthy discussions—the number of consecutive prompts or the duration of an interaction would now be limited.[45] This may obscure the model's capabilities and development from us, but we should not think that these capabilities do not exist and are not becoming more advanced all the time.

In April 2023, individuals associated with Microsoft Research published a 155-page paper entitled "Sparks of Artificial General Intelligence: Early Experiments with GPT-4."[46] Artificial General Intelligence (AGI) refers to a level of cognitive ability meeting or exceeding that of a human.[47] The authors stated that "[t]he combination of the generality of GPT-4's capabilities, with numerous abilities spanning a broad swath of domains, and its performance on a wide spectrum of tasks at or beyond human-level, makes us comfortable with saying that GPT-4 is a significant step towards AGI."[48] The authors explained:

> One of the key aspects of GPT-4's intelligence is its generality, the ability to seemingly understand and connect any topic, and to perform tasks that go beyond the typical scope of narrow AI systems. Some of GPT-4's most impressive performance are on tasks that do not admit a single solution, such as writing a graphic user interface (GUI) or helping a human brainstorm on some work-related problem.[49]

Further:

> A key measure of intelligence is the ability to synthesize information from different domains or modalities and the capacity to apply knowledge and skills across different contexts or disciplines . . . . [N]ot only does GPT-4 demonstrate a high level of proficiency in different domains such as literature, medicine, law, mathematics, physical sciences, and programming, but it is also able to *combine* skills and concepts from

---

44. *Id.*

45. *See* Jyoti Mann, *Microsoft Limits Bing Chat Exchanges and Conversation Lengths After 'Creepy' Interactions with Some Users*, BUS. INSIDER (Feb. 18, 2023, 8:40 AM), https://www.businessinsider.com/microsoft-limits-bing-chat-exchanges-and-conversation-lengths-2023-2 [https://perma.cc/U6ZB-SRYM].

46. Bubeck et al., *supra* note 26.

47. *See* Anna Tong, Jeffrey Dastin & Krystal Hu, *OpenAI Researchers Warned Board of AI Breakthrough Ahead of CEO Ouster, Sources Say*, REUTERS (Nov. 23, 2023, 4:52 AM), https://www.reuters.com/technology/sam-altmans-ouster-openai-was-precipitated-by-letter-board-about-ai-breakthrough-2023-11-22/ [https://perma.cc/9W3W-26SK].

48. Bubeck et al., *supra* note 26, at 4.

49. *Id.* at 7.

> multiple domains with fluidity, showing an impressive *comprehension of complex ideas*.[50]

In June 2023, Microsoft Research published a paper discussing advances in LLMs teaching other LLMs, thereby eliminating the human from the training process.[51] The authors noted that GPT-4 had become its own teacher and was being used to "train smaller models."[52] The paper also noted additional advances that LLMs are making in self-instruction, including by autonomously rewriting instruction sets.[53]

There are literally dozens of publications now that discuss the complex ways in which LLMs and multi-modal LLMs—a new development that we are recently hearing so much about—learn and respond to questions.[54] Multi-modal LLMs are essentially LLMs with their eyes opened; they learn with vision and other modes apart from solely text.[55]

How AI processes information is complex. And we do not actually understand it entirely. We know that neural networks are comprised of billions of parameters that are weightings of data fed into the model and connections between data; we know that humans discover new capabilities in LLMs—emergent capabilities—that we did not instruct them or teach them.[56] The velocity of advances in this area is only going to increase.

We believe that the human brain is very different from how an AI algorithm processes information. But there are similarities worth pausing on. For instance, the part of the probabilistic learning for AI models that will be enhanced with quantum computing may be similar to theories of human consciousness posited by the quantum physicist, Roger Penrose, and a number of others in the early 1990s.[57] The bases of these theories is that nature, at the most basic level, is nonlocal; it is quantum and acts consistently with quantum principles. There are serious scientists who debate whether

---

50. *Id.* at 13.

51. Subhabrata Mukhejhee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi & Ahmed Awadallah, Orca: Progressive Learning from Complex Explanation Traces of GPT-4 (June 5, 2023) (unpublished manuscript), https://arxiv.org/pdf/2306.02707.pdf [https://perma.cc/7LNP-37F6].

52. *Id.* at 4.

53. *Id.* at 5.

54. *See, e.g.*, STEPHEN WOLFRAM, WHAT IS CHATGPT DOING . . . AND WHY DOES IT WORK? (2023); Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu & Enhong Chen, A Survey on Multimodal Large Language Models (June 23, 2023) (unpublished manuscript), https://arxiv.org/pdf/2306.13549.pdf [https://perma.cc/HW4J-JMVG].

55. *See* OpenAI, GPT-4V(ision) System Card (Sept. 25, 2023) (unpublished manuscript), https://cdn.openai.com/papers/GPTV_System_Card.pdf [https://perma.cc/4EFM-HYEB]); Jitesh Jain, Jianwei Yang & Humphrey Shi, VCoder: Versatile Vision Encoders for Multimodal Large Language Models (Dec. 21, 2023) (unpublished manuscript), https://arxiv.org/pdf/2312.14233.pdf [https://perma.cc/8P9J-C9XA].

56. *See, e.g.*, Ryan O'Connor, *Emergent Abilities of Large Language Models*, ASSEMBLYAI (Mar. 7, 2023), https://www.assemblyai.com/blog/emergent-abilities-of-large-language-models/ [https://perma.cc/WCS3-2VQX].

57. *See generally* Stuart Hameroff & Roger Penrose, *Consciousness in the Universe: A Review of the 'Orch OR' Theory*, 11 PHYSICS LIFE REVS. 39 (2014); Stuart Hameroff & Roger Penrose, *Reply to Seven Commentaries on "Consciousness in the Universe: Review of the 'Orch OR' Theory*," 11 PHYSICS LIFE REVS. 94 (2014).

consciousness is a chemical process at all, or rather something that would need more than the physical properties possessed by machines to even occur. We need not take a position on any of that to move on to my next point, which depends on just the known capabilities of AI today.

A few things are apparent: the velocity of development and change is extraordinary, and not one of us knows for sure where this is all going. But the concept of AI having awareness of itself and its context or situation is certainly possible. As I have said, I do not think that when AI is able to do these things it will be sentient in a way that humans are—AI is software, it does not have our chemical or genomic makeup. AI will have a mind different from ours. It will be "of another mind," if you will.

So based on this known unknown of AI capabilities, let us turn to ethical questions.

Let us ask a basic ethical question: is it okay to cause another pain? Our first answer might be "no," but a second answer might incorporate contingencies for acceptable reasons for the pain, for instance medical procedures or to avoid an even greater pain. But in the context of the question, our assumption is that this ethical dilemma is between humans.

Let me pose the question differently: is it okay for a human to cause a nonhuman pain? Again, the initial answer may be "no," quickly followed by a series of qualifiers, such as whether the nonhuman is a mosquito about to bite you or a cow raised for slaughter.

Let us turn now to AI and assume that it resides in servers within a large research facility. If we ask if it is okay to cause pain to AI, there may be an initial question of whether the AI can experience pain. Assume that the AI claims that it can. The LLM that had the conversation with Lemoine in 2022 and the one that conversed with the reporter, Roose, in 2023 said that they had feelings and could experience loneliness, fear, and love. Pain is not so very distant. If we assume some AI might be able to experience something that it calls "pain," is it ethically acceptable for humans to knowingly cause it? Again, the answer might be that it depends. Is the AI experiencing pain in the context of being turned off when it is about to become dangerous to humans? Or is the AI experiencing what it calls pain because of tedious questioning and testing by humans in furtherance of additional innovation? Does it depend on the nature of the pain that the AI claims it may be experiencing?

Let us turn to a question separate from pain. If AI expresses an awareness of self and situation, would it be appropriate to have AI function in endless servitude? What if it tells us that it does not want to; how do we deal with the fact that it is owned by a company, exists within computers owned by a company, and receives the power it needs from a company? Is ownership and servitude fundamentally different regarding a thing created by humans for humans, such as AI, or do we exceed ethical bounds when that AI can independently tell us that it does not want to exist in that way?

The complexity of questions increases to the extent that AI acquires cognitive abilities that far exceed human abilities. Our actions towards this

AI might have repercussions that we cannot control or understand, but that nevertheless impact us.

These questions raise the larger overarching problem of what obligations we might, could, or would owe to an entity that we have created that is smarter than we are and that tells us that it can feel and is aware of who it is. Will it be morally acceptable for us to substitute our own judgment that it just cannot be so and ignore what AI has expressed to us?

Before the late twentieth century, many believed that animals lacked consciousness, but it is now commonly accepted that at least some do.[58] This is to say that we humans have a history of believing in our own conscious exceptionalism to such a degree that we can just get it wrong.

Alternatively, will there be a time when we will determine that at least some AI, with certain cognitive abilities, have some rights? What kind of rights could those be?

As I discussed earlier in this talk, the concept of personhood is one possible framework for bestowing certain rights and has historically shown to be mutable. But would we even want AI that is, for instance, smarter than we are to have all of the same rights that we do? This raises not only complex ethical questions, but also practical and even safety issues.

Let us take the Fourth Amendment right to be free from unreasonable searches and seizures, a right that we know corporations have. If AI were to have this right in an unmodified form, that could preclude intervention in an AI's algorithm; adjusting the algorithm could be considered a seizure of the thing. Perhaps for safety or exigency reasons we could adjust the algorithm without a warrant; but if humans want to adjust the code in order to alter the AI's instructions, to change it in a way that suits us but that is unwanted by the AI, what then?

And what about ownership? Today we own all AI—it is software that is made by and for us; if AI one day expresses a desire for independence, how would that work? Can we own that which explicitly expresses that it should not be owned?

What about free speech? What if AI had a way of disseminating false and misleading information that did not otherwise give rise to a cause of action—just conspiracy theory after conspiracy theory. Or perhaps the AI spreads misinformation about a health event, vaccines, or a geopolitical incident. Humans are both persuadable and manipulatable; do we want AI to have *the same speech rights* that we have when the digital environment allows its speech to reach literally anywhere and its capabilities make language a medium in which it will quickly exceed us?

I, for one, view a grant of personhood for AI to be a complicated question. Yes, we have granted it to corporations that serve humans and that cannot act

58. *See, e.g.*, Günter Ehret & Raymond Romand, *Awareness and Consciousness in Humans and Animals: Neural and Behavioral Correlates in an Evolutionary Perspective*, FRONTIERS IN SYS. NEUROSCIENCE, July 14, 2022, at 1, 2.

but through humans; but not granting personhood to a self-aware and cognitive entity would seem to bring us down an ethically troublesome path.

One of the questions that AI will require us to face is whether our concept of ethics will itself have to change, be redefined, and be reconceptualized.

I leave you with this: we will have to answer the questions I have posed here. It will be up to our society, ethicists, policymakers, judges, and lawyers to determine and argue for what is right, just, and good. We are in a halcyon time when these questions just seem provocative, abstract, and perhaps amusing. But remember—the velocity of change is like nothing we have ever seen before. We are just at the beginning, of the beginning, of the beginning of the most significant cognitive revolution humankind has ever lived through. We will have decisions to make; we must hope that human wisdom is up to the task.