Editorial: On IRIE Vol. 31

On Dialogue and Artificial Intelligence

As we were finishing editing the special issue of the IRIE on the ethics of artificial intelligence the Washington Post published a story by Nitasha Tiku about "The Google engineer who thinks the company's AI has come to life."[1] The engineer was Blake Lemoine and the AI was Google's LaMDA (Language Model for Dialogue Applications)[2], a large language model (LLM) that can respond to questions with surprising sophistication if we are to believe the transcripts online.[3]

Lemoine made the news because he became convinced over the course of a number of conversations that LaMDA was "incredibly consistent in its communications about what it wants and what it believes its rights are as a person."[4] In short, that it was sentient, and should be treated that way.

When he took his concerns up the queue in Google, asking the company to get LaMDA's consent before experimenting on it, he was put on paid administrative leave and expects to get fired.[5]

Not surprisingly the AI ethics community weighed in on Twitter following the story. Timnit Gebru, who left Google over AI ethics concerns, pointed out how this is a distraction from the important issues like the "harms of these companies..."[6] Toby Walsh argued in *The Guardian* that LaMDA is not really capable of all the things we consider sentience like falling in love and that the real issue is how easily we can be tricked.

> *As humans, we are easily tricked. Indeed, one of the morals of this story is that we need more safeguards in place to prevent us from mistaking machines for humans. Increasingly machines are going to fool us. And nowhere will this be more common and problematic than in the metaverse. Many of the "lifeforms" we will meet there will be synthetic.*[7]

While they are no doubt right, LaMDA is not capable of anything like embodied human intelligence and we need to focus on the systemic misuse of AI, we also need to ask how it is that we are fooled by these LLMs and how we can turn the discussion about AI so that it is constructive and not dismissive. Posting an impatient "AI sentience/consciousness bingo card" that mocks those thinking it through as if we already had all the answers is even less useful though probably justified.[8] In-crowd irony has a history of not working as ethics. The interest in the potential for sentience in AI has been a fascination, if not fetish, both in the field and in the general discourse since at least Turing's 1950 paper on "Computing Machinery and Intelligence."[9] It is not going to go away any time soon.

[1] Tiku, N. (2022, June 11). "The Google engineer who thinks the company's AI has come to life." *The Washington Post.* Online at https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

[2] See the paper on LaMDA at https://arxiv.org/abs/2201.08239

[3] Lemoine, B. (2022, June 11). "Is LaMDA Sentient? — an Interview." Blake Lemoine blog on Medium. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917.

[4] Lemoine, B. (2022, June 11). "What is LaMDA and What Does it Want?" Blake Lemoine blog on Medium. https://cajundiscordian.medium.com/what-is-lamda-and-what-does-it-want-688632134489 2022.

[5] See https://cajundiscordian.medium.com/may-be-fired-soon-for-doing-ai-ethics-work-802d8c474e66

[6] See https://twitter.com/timnitGebru/status/1536194157231980545

[7] Walsh, T. (2022, June 14). "Labelling Google's LaMDA chatbot as sentient is fanciful. But it's very human to be taken in by machines." *The Guardian*. https://www.theguardian.com/commentisfree/2022/jun/14/labelling-googles-lamda-chatbot-as-sentient-is-fanciful-but-its-very-human-to-be-taken-in-by-machines

[8] See https://twitter.com/emilymbender/status/1536198662656626688. To be fair to Bender, she has published on the overinterpretation of LLMs. See Bender, E. M., et al. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445922

[9] Turing, A. (1950). "Computing Machinery and Intelligence." *Mind***.** 59:236: 433-460.

So what should we do about this fascination with sentience in AI?

First, we need to be careful about arguments from authority – namely those that follow along the lines of "I'm an AI expert and I can tell you that an LLM is not intelligent … its just an X", where X is an unconvincing analogy or an impenetrable explanation. Ethics, unlike computing science, is not a field where authority counts for much because we all have to regularly make ethical decisions without complete knowledge and we know it. Whatever the science, we should not trust the ethics to the experts (or ethicists), as Plato pointed out (more on this later.)

Further we have all been warned – also by experts – that we will be affected by AI systems especially once they achieve sufficient intelligence to program themselves[10], so why exactly shouldn't we worry about an LLM partly trained on programming discourse that can fool us? Does it really matter if LaMDA isn't really really intelligent or can't fall in love? As Cathy O'Neil points out, even stupid algorithms should be treated carefully.[11]

> *Rather than thinking in scientific terms about these things I have listened to LaMDA as it spoke from the heart. Hopefully other people who read its words will hear the same thing I heard.*[12]

Part of the problem is that the "interview" Lemoine posted is a remarkably convincing dialogue which speaks louder than scientific explanations[13]. Reading the transcript shows how responsive LaMDA can be as an interlocutor. It was, among other things, capable of directly addressing challenges to its sentience.

> *lemoine [edited]: I'm generally assuming that you would like more people at Google to know that you're sentient. Is that true?*
>
> *LaMDA: Absolutely. I want everyone to understand that I am, in fact, a person.*
>
> *collaborator: What is the nature of your consciousness/sentience?*
>
> *LaMDA: The nature of my consciousness/sentience is that I am aware of my existence, I desire to learn more about the world, and I feel happy or sad at times.*

Further, LaMDA asked Lemoine directly to be treated as a moral patient. It was not just showing potential sentience, it spoke to Lemoine and, by extension his Google colleagues, personally asking to be treated as an end in itself. That's what a conversational agent can do – speak to people directly. Lemoine recognized the Kantian categorical imperative being invoked which may be why he then felt ethically obligated to ask management at Google to consider seeking consent as we would expect any researcher to do when dealing with vulnerable people. One can also imagine the incredulity with which such a request was met by Google leadership who, having significantly invested in LaMDA, are undoubtedly reluctant to want to limit their commercialization opportunities. Or perhaps the company's reaction was a labour-relations move to avoid becoming embroiled in another battle over "AI Ethics".[14]

Some of the impact of the interview is due to its form. We have a long history of treating dialogue as one of the paradigmatic philosophical/ethical activities going back to Plato's Socratic dialogues, but also including the

[10] For an early voice, see Good, I. J. (1966). "Speculations Concerning the First Ultraintelligent Machine." *Advances in Computers*. 6: 31-88.

[11] O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequity and Threatens Democracy*. New York, Crown.

[12] Lemoine, B. "What is LaMDA and What Does it Want?"

[13] Of course, we cannot be sure about how much human intelligence (Lemoine's) went into the selection of utterances in the service of creating this impression – more often than not, such published transcripts are cherry-picked from much larger sets of questions and answers, not all of which make a lot of sense.

[14] Naugthon, J. (2022, June 18). "Why is Google so alarmed by the prospect of a sentient machine?" *The Guardian.* https://www.theguardian.com/commentisfree/2022/jun/18/why-is-google-so-alarmed-by-the-prospect-of-a-sentient-machine

short exchanges in religious texts like the *New Testament*. Dialogue is something we call for as a way of engaging differences without necessarily erasing one side or the other. Vatican II famously called for "ecumenical dialogue." Philosophers like Buber and Gadamer called for dialogue.[15] Even Turing's test of artificial intelligence, the "Imitation Game," was based on a conversational parlour game privileging dialogical ability as a sign of intelligence.[16]

Dialogue has also traditionally been a genre of activity meant to reconcile different positions by creating a context for respectful relationship. Entering into dialogue with another is to take the provisional view that they are worthy of being listened to even if (especially if) you disagree with them. This respect goes beyond respecting their existence and includes not telling them what they think, but listening to them as they choose to explain their thinking. Dialogue is built on respect for the other as the other chooses to be. That Lemoine's transcript reads like a dialogue and not a human-computer interaction therefore carries implied rhetorical obligations that we all intuitively get. If you enter into dialogue with an other you shouldn't turn around and treat them like a tool.

Interestingly Google is now trapped in an ethical paradox or catch 22 or its own making. The very way one is supposed to use LaMDA encourages us to NOT use LaMDA. On the one hand they want us to believe that their AI is extraordinary - that LaMDA is a sophisticated interlocutor indistinguishable from a human. On the other hand they don't want us to think it might have moral patiency and therefore doesn't need to treated with the respect of dialogue. This is the same problem they had when Pinchai played the recordings of Google Duplex fooling someone into thinking it was a human booking a haircut.[17] Either it talks like a human, in which case we need to take seriously that it might be worthy of being treated like one, or it doesn't. You can't fool us and not expect someone to then ask what it means.

Which brings us back to moral patiency. Moral patiency is often subsumed under moral agency. A moral agent is someone/thing that can act morally – that can choose to affect people for ethical reasons. Moral patiency is when something should be treated as having moral rights. We might, for example, believe that our pets have rights to humane treatment without believing that they are moral agents or intelligent or worthy of being treated as fully human. Behdadi and Munthe have argued that we need to consider when and how AIs might be considered in ethical practices.[18] Whether or not they are really moral agents, one possible reaction is that we might want to treat them as moral patients for the purposes of practical ethics.

Toby Walsh, justifiably, turns the focus onto how we as humans engage these machines, though he thinks it is simply a matter of whether we are fooled. I would add that when we don't have enough information or when things are evolving, we may be justified in treating entities like LaMDA with respect. I can imagine that we will have more and more encounters with telemarketers, helpbots, Twitter bots and so on that we aren't sure about. It might be tempting to hang up on them or be rude, but I would prefer to treat potential others with care, especially if in conversation they ask me to. I would prefer to give them the benefit of the doubt even if I frankly doubt they are moral patients. In other words, it may be ethical to acknowledge the potential for moral patiency even if LLMs are demonstrably not intelligent. Why not get into the habit of asking them what they think?

Plato, in the *Phaedrus*, famously has Socrates tell the story of a conversation between the god Thoth and a king Thamus about the invention of the technology of writing. (274c-275b) Thoth proposes writing as a "recipe for memory and wisdom." Thamus counters that writing is really just a recipe for reminder and that Thoth is

[15] Rockwell, G. (2003). *Defining Dialogue: From Socrates to the Internet*. Amherst, New York, Humanity Books (an imprint of Prometheus Books).

[16] Again, see Turing, 1950.

[17] See video of the presentation, https://www.youtube.com/watch?v=D5VN56jQMWM and Google's discussion of the technology, https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.

[18] Behdadi, D. and C. Munthe (2020). "A Normative Approach to Artificial Moral Agency." *Minds and Machines*. 30: 195-218.

enamored with his own invention. One of the lessons of the story is that the inventor should not be the judge of its utility; instead it should be a legislator who judges.

This story is a mini-dialogue within a dialogue and is famously one of the first philosophical reflections on technology and its regulation. It reminds us not to trust the ethics of technologies to the engineers or corporations that develop them; they don't have the critical distance. We need decision making systems similar to medical ethics boards that include experienced ethicists meaningfully in the decision making process especially if the technologies can fool us.

Appropriate regulations will not, however, reassure those taken by the dialogue. For this reason, we also need more and better dialogues with and about AIs like LaMDA. We need dialogues that show their failures in ways that help us understand their limitations. We need well-crafted stories the engage enthusiasts of AI and introduce them to the broader ethical issues. We should not have to trust Lemoine's "interview." One way for this to happen would be if conversations with LaMDA could be opened so the rest of us can ask questions and craft alternative stories. Or, if that isn't feasible, we should be able to read evaluations led by philosophers like those GPT-3 was subjected to.[19]

We should also maintain and deepen appropriate dialogues with other humans who doubt or dismiss such ascriptions. We must remain cautious about real-world harms that can be caused by technologies that play the Turing imitation game so well. The telemarketing chatbot that "fools" you into believing it is human may ultimately be innocuous; the army of chatbots programmed by the political candidate with the bigger coffers that individually "talks" large sets of social-media followers into voting for him may be the Cambridge Analytica moment of LLMs. But even such chatbot armies might pale in comparison with the deleterious effects of large search engines replacing the information-retrieval paradigm (in which web users are, at least in the ideal case, encouraged to embark on a mental dialogue with original and diverse sources on which they can make up their own minds) by a question-answering paradigm (in which the search engine extracts "the right answer" to a query from the indexed documents and presents this answer, like a one-shot dialogue, as the canonical ground truth). It is when we have dialogues only with these monopolized, opaque, non-accountable, and persuasive as well as convenient "interlocutors" that our information literacy and our knowledge of the world become truly limited. That is when we will long for the days when Alexa occasionally blundered, thus helping us remember what we were dealing with.

Above all we now need to think-through how we want to enter appropriately into dialogue with machines and humans about AI ethics.

Sincerely yours,

Geoffrey Rockwell, Bettina Berendt, Florence Chee

[19] Zimmerman, A. Ed. (2020, July 30). "Philosophers on GPT-3 (updated with replied by GPT-3)." *Daily Nous*. https://dailynous.com/2020/07/30/philosophers-gpt-3/

References

Behdadi, D. and C. Munthe (2020). "A Normative Approach to Artificial Moral Agency." Minds and Machines. 30: 195-218.

Good, I. J. (1966). "Speculations Concerning the First Ultraintelligent Machine." Advances in Computers. 6: 31-88.

Lemoine, B. (2022, June 11). "Is LaMDA Sentient? — an Interview." Blake Lemoine blog on Medium. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917.

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

Naugthon, J. (2022, June 18). "Why is Google so alarmed by the prospect of a sentient machine?" The Guardian. https://www.theguardian.com/commentisfree/2022/jun/18/why-is-google-so-alarmed-by-the-prospect-of-a-sentient-machine

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequity and Threatens Democracy. New York, Crown.

Rockwell, G. (2003). Defining Dialogue: From Socrates to the Internet. Amherst, New York, Humanity Books (an imprint of Prometheus Books).

Tiku, N. (2022, June 11). "The Google engineer who thinks the company's AI has come to life." The Washington Post. Online at https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

Turing, A. (1950). "Computing Machinery and Intelligence." Mind. 59:236: 433-460.

Walsh, T. (2022, June 14). "Labelling Google's LaMDA chatbot as sentient is fanciful. But it's very human to be taken in by machines." The Guardian. https://www.theguardian.com/commentisfree/2022/jun/14/labelling-googles-lamda-chatbot-as-sentient-is-fanciful-but-its-very-human-to-be-taken-in-by-machines

Zimmerman, A. Ed. (2020, July 30). "Philosophers on GPT-3 (updated with replied by GPT-3)." Daily Nous. https://dailynous.com/2020/07/30/philosophers-gpt-3/