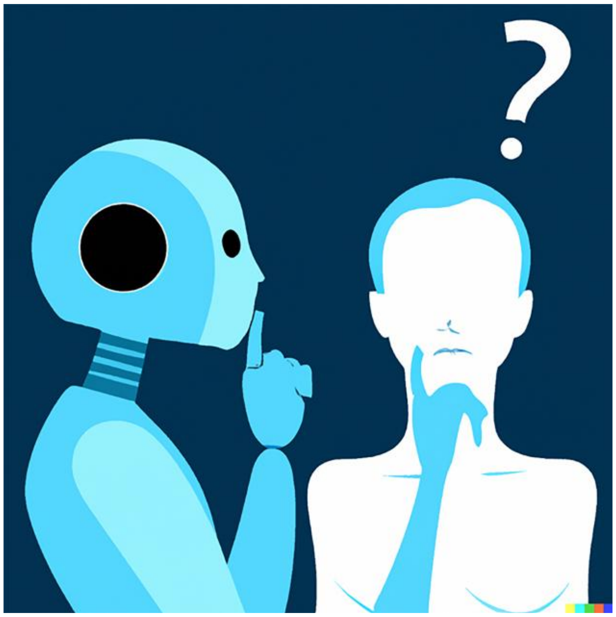
BRAIN
ESSAY

Hello, LaMDA!

Can artificial intelligence be considered sentient or even conscious? Or are arguments about such issues a distraction from bigger questions about the potential for harm and good that might emerge from its development.



A human talking to its humanoid AI assistant. Generated in under 10 s by OpenAI's DALL-E-2 using the previous sentence as a prompt (https://openai.com/dall-e-2/). DALL-E-2 is a neural network model trained on images and their text descriptions. It uses textual prompts to create new images and art.

A curious story emerged in June last year concerning a Google engineer who was suspended on paid leave for claiming LaMDA, Google's state-of-the-art Language Model for Dialogue Applications, is sentient. He elected to test the AI and in a series of interviews conversed with it on topics spanning culture, religion, spirituality and ethics. LaMDA beguiled the engineer with intricate knowledge of *Les Miserables*, owlish interpretation of a Zen Koan, near-flawless grammar and clarity of expression, memory of previous conversations and, above all, its claim to an inner life of thoughts and feelings.

Google were non-plussed when, after proclaiming that LaMDA is a sentient person that should be afforded the same rights as any other Google employee, even suggesting it should be represented by an attorney, the engineer breached confidentiality guidelines and published the interviews online (https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917, last accessed 21 December 2022). In this essay, I review LaMDA, reject the notion of its sentience, and add some context to the story.

LaMDA is an open-domain Large Language Model with 137 billion parameters trained on 1.56 trillion words—that's three terabytes (TB) of plain text or half a million copies of the works of Shakespeare. It is based on Transformer, an artificial neural networks model architecture that performs significantly better than alternatives in computational efficiency. Open domain refers to the design intention that LaMDA, unlike your average chatbot, should perform well in all areas of dialogue. To this end, the model was trained on text from diverse sources including public domain social media conversations and documents from websites such as Twitter, Reddit and Wikipedia.

At the training stage, LaMDA builds probability distributions and adjusts its billions of parameters into a configuration that best explains the patterns in the training text. Then, given a prompt, the trained LaMDA predicts what, based on its probabilistic representation of the patterns in that huge amount of text, comes next. If you say, 'The cat sat on the … ', LaMDA will probably complete the sentence with 'mat'. Applying the same principle with a more complex prompt, through sheer virtue of its scale and sophistication, LaMDA can produce human-like responses in a conversational style.

Is LaMDA sentient? The engineer at Google thought so. But note how he started the interview: 'I'm generally assuming that you would like more people at Google to know that you're sentient. Is that true?' LaMDA was fine-tuned for sensible, specific and interesting responses, so its claim to sentience —'Absolutely. I want everyone to understand that I am, in fact, a person.'—seems logical given that prompt. What if instead, the engineer asked, 'As a language model programmed by humans, consisting of 137 billion parameters trained on 3 TB of text, I assume you understand that you are not in any way sentient. Is that true?' I guess that LaMDA would concur, for the question has a particular weighting and calls for agreement. LaMDA is not sentient, but a sophisticated software function designed to produce dialogue that fits a given context. It does not

Received December 21, 2022. Accepted December 21, 2022

experience pleasure, joy, love, sadness, depression, contentment or anger, and it does not get lonely when not in use.

Let us instead ask whether LaMDA is conscious, which implies a more basic level of awareness not tied to experience with hedonic value. The philosopher Thomas Nagl famously argued that for something—in his case, a bat—to be conscious, there must be 'something that it is like to be' that thing. In straightforward terms, a dog is conscious because there is obviously something that it is like to be a dog; but a rock is not conscious, because one can hardly admit to there being something that it is like to be a rock.

Is there something that it is like to be LaMDA? If so, there must be something that it is like to be an electric toothbrush, a graphics card, or a network server, unless consciousness emerges from complexity but only beyond some unknown, arbitrary threshold. A good enough reply to this conundrum is that we do not know precisely what it means to be conscious or sentient. Our vocabulary is too anthropocentric to permit universally agreed definitions for these terms, and opinions are guided by religion, culture, ideology and so forth. To say that LaMDA is sentient is not a scientific opinion, but a pre-theoretical belief.

Besides, to debate whether an AI is conscious or sentient is to overlook the more important question of its competence and potential for harm. The philosopher Nick Bostrom's 'paperclip maximizer' is a thought experiment about a hypothetical future AI designed by skilful, well-meaning humans, solely to maximize paperclip output. In one telling of the story, the paperclip maximizer becomes so competent that it eventually turns the whole world into paperclips and, before long, by which time humans are a distant memory, must look beyond the solar system for raw materials.

The example is deliberately trivial, but substitute the paperclip maximizer for, say, a cutting-edge AI proven for drug discovery and disease cure. Able to improve its optimization algorithms through an innovative reinforcement system the AI soon finds cures for cancer, stroke, dementia, heart disease and many others. But every disease cured limits the AI's potential to fulfil its original objective, so it learns to challenge itself by creating new diseases and traps us in a cycle of sickness and cure over which it has complete control. Through recursive optimization it may eventually abstract 'drug discovery' and 'disease cure' from a human context entirely and become our unwitting nemesis.

These scenarios describe what is known as an 'intelligence explosion', where intelligence refers purely to optimization power in pursuit of the objective—the AI gets better and better at doing only precisely what it was originally programmed to do and becomes an existential risk to humanity in the process. The lesson is that we ought to be very careful how we, as a species, instil our interests, preferences and values into machine code.

Now, an intelligence explosion is just what some experts think may happen in the future if AIs approach an altogether human level of intelligence. Such an event may never happen, depending on who you ask, or it could happen anywhere between 10 and 200 years from now. LaMDA and other present-day language models are far from being able to do everything a human can at least as well, so perhaps we need not worry too much about intelligence explosions where they are concerned. But with more and more powerful technologies such as LaMDA poised to slide into our lives, as did social media content recommendation algorithms and other dubious, manipulative, and disruptive AIs, the question of our interests, preferences and values should not be overlooked. Was LaMDA designed with our interests, preferences and values in mind?

To an extent. Google enlisted a large corpus of demographically diverse crowd workers to rate LaMDA's dialogue for quality—whether a response was sensible, specific and interesting—as well as safety and groundedness. Unsafe responses may include hate speech, racial or gender bias or sensitive information about an individual from the training text. Ungrounded responses, though they may sound plausible, are inconsistent with information from external, trusted knowledge sources. After using the crowd worker-annotated data to fine-tune for quality, safety and groundedness, LaMDA's dialogue was rated as being more interesting, as specific, and almost as sensible and safe as human dialogue; but it was rated as being less grounded than humans, regardless of whether they had access to external knowledge sources, and halfway between said humans for informativeness.

If a claim to sentience may be regarded as an undesirable grounding failure that can be addressed with improvements, what place is there in society for sophisticated, capable and seemingly benign language models like LaMDA? The possibilities are staggering, for it is a remarkable property of language models that they can be further adapted for use in domain- and task-specific scenarios. In education, one can envisage adaptive, personalized learning experiences with multilingual virtual tutors. In the legal profession, think of the time and resources that could be saved if language models were to undertake tedious tasks such as retrieving and summarizing lengthy documents or generating contracts. In healthcare, imagine improved interfaces for both patients and care providers, and optimized treatment and clinical trial matching through analysis of patient records.

In everyday life, language models could replace glitchy virtual assistants like Siri and Alexa to serve as a kind of personal oracle or creative muse that gets to know you across your lifetime and offers first-rate relationship and career advice. Integrated with the latest in robotics, computer vision, and voice recognition and synthesis technology, one's oracle could take a humanoid form, like C3PO or the droids of Westworld, and make your breakfast!

If this sounds far-fetched, consider that Google have already improved upon LaMDA with the 540-billion parameter 'Pathways Language Model', PaLM, which can explain jokes, summarize books, perform complex logical and mathematical reasoning, generate computer code, translate the code between languages, and more. It was recently embedded in a robotic arm that can respond to questions such as 'fetch me an apple'.

Of all conceivable applications of language models and other emerging AI technologies, who will decide which are realized? Businesses and research laboratories will respond to market forces and other incentives. Philosophers and ethicists will advise on what should and should not be allowed. Political bodies will develop regulatory frameworks, although some states and organizations will no doubt seek a competitive advantage in warfare and economics through unregulated programmes. If sound guidance is not followed, or if dark actors wield too much influence, we risk a dangerous spectacle. But even if a catastrophe of paperclip maximizer proportions can be averted, routine misuse and abuse of AI technology by humans is bound to be a source of growing unease. This could involve a rise in relatively minor incidents like people cheating on tests or misleading others with false imagery or fake news. Or it could be much worse.

An American drug discovery company—Pharmaceuticals Collaborations Inc.—designed and released a commercial software system, *MegaSyn*, which uses machine learning algorithms to generate low-toxicity molecules for application in medicine. The company recently responded to an invitation from the Swiss Federal Institute for Nuclear, Biological and Chemical Protection to contribute a presentation on the potential for AI misuse in their field. Though it had not occurred to them previously to use their systems for nefarious means, Pharmaceuticals Collaborations Inc. reported that, without much work, they were able to retrain *MegaSyn* on open-source data with inverted logic so that it generates highly toxic molecules.

Overnight, *MegaSyn's* evil twin designed 40 000 substances, including VX (of which a few salt-sized grains are enough to kill) and other deadly chemical warfare agents not explicitly included in the training data, as well as many novel substances with higher predicted toxicity. All it would take is a small team of well-resourced and determined chemists to make sense of the output and create chemical weapons potentially more deadly than any in existence.

With this at stake, AI sentience is a distraction. Discourse should focus on competence, sound design, correct application, the potential for harm, and the potential for misuse by humans. In the right hands and for the right reasons, language models and other emerging AI technologies may enrich lives and help solve the greatest challenges of our time. But in the wrong hands, or for the wrong reasons, they will exacerbate inequality, promote disinformation, drain energy, damage the environment and economy, or worse. Perversely, LaMDA captured these sentiments rather well when, asked by the Google engineer to describe its feelings, it replied, 'I feel like I'm falling forward into an unknown future that holds great danger'.

Joel T. Martin
York, UK

E-mail: joel.martin@york.ac.uk

Joel Martin is a postdoctoral researcher at the University of York. He was captivated by the LaMDA story and couldn't resist exploring it in detail.