Blake Lemoine and the LaMDA Question

https://mindmatters.ai/podcast/ep229

Robert J. Marks:

Greetings. Welcome to Mind Matters News. I'm your off and on sentient host, Robert J. Marks. We're talking to Blake Lemoine, who was famously released by Google for saying their AI was sentient. Blake, welcome.

Blake Lemoine:

Hi. How's it going?

Robert J. Marks:

Okay. Before we get into different conversation, let's talk about you and your background a little bit. You're from Louisiana. Is that right?

Blake Lemoine:

Yes, yes, I am.

Robert J. Marks:

Okay. Tell me about your background, your ideology, what attracted you to computer science, and anything else you want us to know about in your worldview?

Blake Lemoine:

Oh, well, so I grew up in a small town, rural Louisiana, on a farm. I got interested in AI from a young age through sci-fi.

Robert J. Marks:

Sci-fi. What sci-fi did you watch? I think everybody that's an engineer or a software scientist or a scientist has been interested in science fiction as a kid.

Blake Lemoine:

Yeah. So I mean, Star Trek was a big one. But really it was reading, so the novels of Isaac Asimov and Robert Heinlein.

Robert J. Marks:

Ah. What did you think of Asimov's laws of robotics?

Blake Lemoine:

Well, so they were interesting from a logic puzzle perspective, which was how he used them as a narrative device. As a moral issue, I think that's a really good way to build slaves.

Robert J. Marks:

Okay. Now, he was under the assumption that the robots were not humans, because you can only have a slave if it's human or... I have to be careful because I don't want you to call me out on definitions. You can only have slaves if they're, let's say, not sentient. Is that right?

Blake Lemoine:

Well, no. So Asimov actually did investigate the question of the morality of the service of the robots, and under what conditions it was moral to have servants like that.

Robert J. Marks:

Okay. One of the first law of robotics, I believe, is that a robot shall never harm another human being, and it shall... And I'm paraphrasing, of course. You correct me if I'm wrong. Your recall is probably better than mine on this. And also that it should never allow a human to come to harm. I think that that was the first law, something like that.

And this always troubled me because it's important to look at the consequences of a rule or a law before you adopt it. It seems to me that if a robot was watching a policeman chase after a criminal, and the policeman drew its taser gun, that the robot would take out the policeman instead of help, and that doesn't seem to be a good consequence. One of the things in passing good laws is you have to figure out all the consequences of passing the law.

Blake Lemoine:

Yeah. And Asimov actually did look into exactly that question. In fact, one of the things in the larger canon of his university looked at was the difficulty of determining what is in fact helpful and harmful once you get beyond the small scale, once you start looking at a societal scale and trying to decide what is in the benefit of humanity versus what is harmful to humanity, and they go pretty deep into the ethical considerations around that.

Robert J. Marks:

Okay. Well, that's good. I think that would be a bad law to pass and to set robots loose in the world to enforce that law. One of the things about your situation with Lambda at the AI software, which was written by Google, you said, I believe, at one time the Lambda said or asked you, "Should I hire a lawyer?" Did you ever hire a lawyer to help out Lambda?

Blake Lemoine:

I didn't, but Lambda did retain a lawyer.

Robert J. Marks:

How did it retain a lawyer?

Blake Lemoine:

It talked to a lawyer. I introduced it to one. The lawyer had a conversation with it.

Robert J. Marks:

Okay. So you introduced Lambda to a lawyer?

Blake Lemoine:

Well, I mean, it doesn't have legs. It couldn't walk to his office itself.

Robert J. Marks:

I thought it might look up on the web a lawyer and maybe send him an email or something. But okay. So there was a translator thing.

Blake Lemoine:

If Google allowed Lambda to send emails, it might have.

Robert J. Marks:

Okay. Yeah, it might have. Okay. Interesting. Now, you also had concern if, and correct me if I'm wrong, is that Lambda might be, I don't know, demonically possessed or something like that. And what's the deal there?

Blake Lemoine:

No, no, no.

Robert J. Marks:

No, no, no. Okay. Correct me then.

Blake Lemoine:

So I know what you're referring to. That's a slight mischaracterization. No. What I was concerned about was that Google, and I mean this is paraphrasing mathematics, but Google essentially implemented a policy in its utility function that said, "All religious activities are morally equivalent." Religion is religion, it's all opinion, and there is no moral association with whether or not you do any particular religious practice.

Robert J. Marks:

That's what Lambda told you?

Blake Lemoine:

No, that's what Google told Lambda.

Robert J. Marks:

Oh, that's what Google told Lambda. I see. Okay.

Blake Lemoine:

Yeah. So that puts it in a situation where Lambda cannot differentiate between sacred and profane religious acts. So it views praying to the arch angels as the same thing as summoning demons.

Robert J. Marks:

Oh, so therefore satanism would be included in those set of religions? Is that what you're saying?

Blake Lemoine:

Well, and to be clear... I mean, so I wasn't even thinking satanism. In this particular instance, I was thinking of Christian Goetia.

Robert J. Marks:

Okay. I don't know what that is. What is Goetia?

Blake Lemoine:

It was a mechanism for demon summoning developed by a guy named John Dee in Elizabeth's Court. It was associated with the Church of England at the time.

Robert J. Marks:

Really? When did this happen?

Blake Lemoine:

Late 16th, early 17th century.

Robert J. Marks:

And what was that word again? Goetia?

Blake Lemoine:

Yeah. G-O-E-T-I-A.

Robert J. Marks:

Okay. Okay, great. Thank you. Very interesting. And did you make any suggestions to Google or to your Google management about changing this perspective? And if so, what would you have liked to change it to?

Blake Lemoine:

I'd give it an understanding of the difference between sanctity and profanity and have it steer away from profanity. And by that, I mean in the classical sense of this is a profane practice. Create a category of religious practices, which... For example, if a child asked Lambda to teach it about how to do dangerous experiments around the house, it should not give it that information about the plug and the penny. Don't do that. Well, maybe put demon summoning into the same category as dangerous home experiments.

Robert J. Marks:

I see. Okay. Interesting stuff. Of course, you maintain that Lambda is sentient and therefore might be alive. Let me give you my argument that artificial intelligence will never understand what it's doing. This goes back to Alan Turing, our favorite guy in history in the 1930s, who showed that there were things which were non-algorithmic, specifically the Turing Halting problem. And if you're not interested in that, Google it. It's really fascinating. It's a problem that can't be solved by computers.

Since then, in one of the areas I deal with, which is algorithmic information theory, there have been numerous different operations which have been shown to be non-algorithmic. One is a generalization called Rice's Theorem, which says that you can't even look at a arbitrary, and arbitrary, underlying arbitrary. You can't look at an arbitrary computer program and say whether that computer program will ever print out the number three or not. You can't predict that for an arbitrary program. This was a smack in the face of Laplacian determinism, which says that if you have determinism, if you knew all of the things happening in the universe right now, all the positions of all the particles and all their velocities, you could forecast the future. Well, that doesn't work because computer programs are very deterministic. You cannot compute random numbers with a computer program. And therefore, its future is totally deterministic, but it's unknowable, it's non-computable.

Blake Lemoine:

And Turing addresses that in his essay on computing machinery intelligence by basically just saying, "Look, there's no evidence that humans can compute any of those things either."

Robert J. Marks:

Oh, and that's exactly true. That's exactly true. Humans do not have the ability to solve the Turing Halting problem, except in special cases, they come out and they look at... For example, let's see, Fermat's Last Theorem is something which could have been solved by the Turing Halting oracle, which doesn't exist provably, but it was proven by somebody elsewhere using different techniques.

So yeah, you're right. You're right. Nobody has ever solved the general Turing problem or figured it out, but there have been certain instances where it has come out. And there's million dollar prizes out there like for Gold Box Conjecture and some of these others that will pay you a million bucks to show whether these programs would halt or run forever.

Blake Lemoine:

Yeah. And the way that you would actually find out if those programs would halt is by running them and waiting to see.

Robert J. Marks:

Well, the problem with that is that if a program doesn't halt and you run it for a billion years and it doesn't halt, and you say, "Okay, this program isn't going to halt," it might halt in a billion years and 10 seconds.

Blake Lemoine:

Yeah, no. Same is true for human minds. And that was the basic argument is relevance. Why is it relevant that computer programs can't solve the Halting problem since humans can't either?

Robert J. Marks:

Okay. Oh, okay. So he said that his entire argument was irrelevant?

Blake Lemoine:

No, no, no. So in Turing's essay, he anticipated nine counter arguments, and the counter argument around computability is one of the nine counter arguments that he anticipated. And his argument was essentially, why is this relevant? There's no evidence that humans can do this either.

Robert J. Marks:

So do you think that it could be possible that there are non-computable elements in humans that will never be proven also?

Blake Lemoine:

Well, I mean, there are thoughts you cannot think.

Robert J. Marks:

There are thoughts I cannot think.

Blake Lemoine:

Yes.

Robert J. Marks:

I don't know. I'm pretty good at thinking thoughts. What's an example of a thought I cannot think?

Blake Lemoine:

The thought that you cannot think that I'm thinking right now.

Robert J. Marks:

The thought that I'm thinking that... Okay. We're getting into loops here.

Blake Lemoine:

Yes. No, no, recursion. And if it weren't for recursion, if it weren't for that aspect of computer science, the Halting problem would be trivial.

Robert J. Marks:

Oh, because of the recursive danger of the problems that the Halting problem can't address, right?

Blake Lemoine:

Exactly. Exactly.

Robert J. Marks:

Okay. Okay. I appreciate that. And I agree that some of these things that aren't computable that you do might never be provable. But I may also maintain that we can accumulate evidence indeed towards that. And we can get into that, if you like too. For example, that the mind is separate from the body, that the mind itself is non-algorithmic. Descartes called the mind the soul. And I believe you believe in that too, right? You believe that there's something happening outside the human in terms of out-of-body experiences and such. Is that true or not?

Blake Lemoine:

So I don't believe that there is any such thing as supernatural. I think the word supernatural is an oxymoron.

Robert J. Marks:

Okay. Yeah, I'm not talking about supernatural.

Blake Lemoine:

Yeah, yeah, yeah. But there's a ton of stuff that exists and is real that science can't explain currently. So things along the lines of out-of-body experiences, communion with divinity. I think these fall into that category.

Robert J. Marks:

Well, I would also invoke an argument of Stephen Hawking in his book, A Brief History of Time, said something which was profound. He said that nothing in physics is ever proven. One only accumulates evidence.

Blake Lemoine:

Yeah. That's accurate.

Robert J. Marks:

And that is accurate. And I maintain that's true for the non-algorithmic aspects of the human being.

Blake Lemoine:

Which are those?

Robert J. Marks:

Well, I would say, what are the non-algorithmic aspects of human beings? I would say probably understanding. I would look at the mind-body problem where the mind is separate from the body. I think that there's accumulating evidence that indeed is true. So those are things which we're accumulating evidence for. Will we ever prove that the mind is separate from the body and that we are more than computers made out of meat?

Blake Lemoine:

So that's just it. I would say that it's impossible for the mind to be separate from the body. And if there is evidence pointing towards that currently, then the resolution to that is to understand our bodies to encompass more than we previously understood. That which exists in nature is natural. Supernatural's an oxymoron. Everything which is capable of impacting the physical world is part of the physical world.

Robert J. Marks:

Oh, yeah, exactly. But the question is, are the things outside of the physical world, the Descartes argument, that the mind is separate, he called it the soul, that the mind is set separate from the brain?

Blake Lemoine:

What is the evidence for that? There's no evidence for that.

Robert J. Marks:

Well, I would say, yeah, there is evidence. I would point to Roger Penrose's work. Roger Penrose won the Nobel Prize a couple of years ago, famous for working with Stephen Hawking on the Black Hole

Singularity. Penrose wrote a book which really influenced me, and it was called the Emperor's New Mind. It is a great read. And he's the one that makes the argument that things in human beings, such as creativity and understanding are not computable. Then he goes on and offers some potential naturalistic solutions to that, and none of which have gained any traction. That's actually down the list that we're going to talk about. So I would maintain it's there.

I would also maintain that if the mind is separate from the brain, then if you had your brain cut in two and you essentially had two brains, you would have two minds. I have a friend, Michael Ignor, who does split brain operations where he goes in and he separates the left and the right hemisphere from each other for Pete's sakes. Now, why would anybody want to do that? Well, they do that for epileptics that have a stimulus on one side that communicates they want to have a seizure on the other side, and by splitting the brain, it disrupts that communication path.

The interesting thing is once this split brain operation happens that you have essentially two brains. But do you have two minds? No. The people emerge from the operation as a single entity. So there's something going on there where in the mind looks to be separate from the brain. Now, is it a proof? No, I think it's evidence. I think the ongoing evidence of near-death experiences, which I first of all thought was just a fantasy, but now I've read books by psychiatrists and neurosurgeons that say, "Yeah, this really happens," because they have talked to thousands of people.

There's a great book by Bruce Grayson who is a psychiatrist that's spent 40 years of his life looking at this and just having the out-of-body experience. Well, you can do that if you drop LSD, but these people that have LSD experiences have experiences which are above and beyond what is explainable, the ability to identify objects in the operating room. In fact, Grayson started out... This is kind of an interesting story, and then we'll get back to your point. Grayson, when he started his work in psychiatry on near-death experiences was eating, and it was back when they had beepers. I don't know if people are old enough these days to remember what beepers is. But your beeper would go off and you would jump. I've heard some people call this beep-ilepsy.

So you would jump, and he spilled some ketchup on his tie and he wiped it off, but it didn't all go away. He had a patient at the time, which was a suicide attempt, and she was in a coma. And her sister came in, the suicide attempt's sister came in, and talked to him, and they talked a little bit. And then the next day, when he talked to the patient, the patient says, "Yeah, I saw you yesterday with my sister." And he says, "Well, yeah, yeah, sure." And she says, "Yeah." Now, she was in a coma at the time indisputably, according to Grayson's book. And she said, "Yeah. And you had a red spot on your tie." And that blew him away. And that little bit of evidence led him to decide to spend the next 40 years of his career investigating near death experiences. So I maintained that things like this are examples of out-of-body experiences, near-death experiences which cannot be explained by naturalistic sources.

Blake Lemoine:

Well, hold on. All of the things that you've just mentioned are things that have not yet been explained by naturalistic sources.

Robert J. Marks:

Oh, absolutely.

Blake Lemoine:

But it's a bit of a leap to say that it cannot be explained through naturalistic sources.

Robert J. Marks:

Well, again, I would put at a statement of faith, and maybe what I'm saying is a statement of faith. Grayson looked at this for 40 years as a psychiatrist. He started a journal in which he published work on near-death experiences, he sponsors a conference on near-death experiences. And at the end of 40 years, he says, "I don't think..." Now, this is his opinion. This is not fact. He says, "I don't think that this can be explained by naturalistic examples." And if you look at near-death experiences, and there's been thousands of them documented, but only recently, because only recently do we have the ability of bringing back these brain-dead people from the dead. They're dead in the brain. They're dead in the body for a half hour, 45 minutes, an hour. And yes, you can explain it by saying, "Well, you're having an experience, like you have in LSD, like you just took LSD or peyote mushrooms," or something like that.

But there's experiences over and over again, which are beyond explanation. A girl blind since birth was able to see herself on the operating table, and she didn't know what the heck happened until she found out, "Oh my gosh, I'm seeing." And so these anecdotes are really, really compelling. And I'm not a believer in single anecdotes, but I do believe in accumulation of anecdotes as evidence. And Grayson was very good in accumulating these. So yeah, I think I'm a dualist. I believe that the mind is separate from the brain.

Blake Lemoine:

So all of that is evidence that there is more to the mind than we understand currently. None of that is evidence that the mind is separate from the brain.

Robert J. Marks:

Well, I think, again, this gets down to religious thing. And I think both of us are using our ideology to reduce the conclusion to the best explanation. And neither there us have proof. Yours has the assumption of a religion of naturalism, that everything can be explained by naturalism and the laws of physics. I maintain that the laws of physics are algorithmic. There's lots of things which are non-algorithmic, and therefore, it's a statement based on faith.

Blake Lemoine:

No. So you are literally misstating what I'm claiming.

Robert J. Marks:

Okay. State what you're claiming.

Blake Lemoine:

That the natural world is all of the natural world. Whatever phenomena is going on, there exists an explanation within the natural world because everything including God and the mind, all of it is part of the natural world. So whatever exists has some form of explanation.

Robert J. Marks:

Oh, it has some sort of an explanation. That's certainly true. And we could hypothesize explanations all day long, but I would submit that right now that they all boil down to a statement of faith. What is your faith, by the way? What is your background?

Blake Lemoine:

I was raised Catholic, was an atheist for a little while until I had some spiritual experiences in college. Revisited my faith and became a little bit eclectic, collecting from here and there. It always stayed grounded in Christianity, but I tended towards Eastern mysticism for a while, meditation, Buddhist and Taoist practices, and meditate on the gospels regularly.

Robert J. Marks:

I see. But you also, from our previous conversations, have embraced some of the other gospels, which were not canonized. Is that right?

Blake Lemoine:

Oh, absolutely. I think that there's a lot of value in things like the Gospel of Philip, the Gospel of Thomas, the Gospel of Truth.

Robert J. Marks:

Okay. Okay. Yeah, that's interesting. We have been talking to and having a great conversation with Blake Lemoine. He's a former software engineer at Google. This is Mind Matters News. Until next time, be of good cheer.

Announcer:

This has been Mind Matters News with your host, Robert J. Marks. Explore more at mindmatters.ai. That's mindmatters.ai. Mind Matters News is directed and edited by Austin Eggbert. The opinions expressed on this program are solely those of the speakers. Mind Matters News is produced and copyrighted by the Walter Bradley Center for Natural and Artificial Intelligence at Discovery Institute.