Can AI Ever Be Sentient? A Conversation with Blake Lemoine
https://mindmatters.ai/podcast/ep280

Announcer:

Greetings. Is Google's AI system LaMDA sentient? This was the position held by former Google engineer Blake Lemoine that ultimately culminated in his dismissal from Google. But was he correct? We returned to our discussion with Blake on this topic today on Mind Matters News.

Robert J. Marks:

Greetings. Welcome to Mind Matters News. I'm your sentient host, Robert J. Marks. Our guest today is Blake Lemoine. Blake Lemoine made national headlines when, as an employee of Google, he claimed that Google AI software that was named or dubbed LaMDA was sentient. I met Blake at a recent COSM conference sponsored by the Bradley Center, the same organization that sponsors this podcast. He's well-spoken, he's articulate, and he's a gentleman and I'm glad that I know him.

Blake is famous. Because of his actions at Google, he has been interviewed on such platforms as Bloomberg News, and he was even interviewed by Tucker Carlson on Fox News. Both Blake and I agree that Google's LaMDA is an amazing piece of software. It could do incredible things, but our viewpoints differ widely on whether LaMDA is sentient like a human. Blake believes that LaMDA might be sentient. I maintain that AI can never understand what it's doing, will never be creative, nor be sentient or conscious. AI can mimic sentience, I believe, but never duplicate it. And we're going to be having a back and forth on this in the podcast.

Let me give you a little background. Blake earned his master's in computer science from the University of Louisiana in Lafayette in 2010, and went immediately to work at Google. So Blake, welcome.

Blake Lemoine:

That's actually not accurate. I spent some more time working on a PhD.

Robert J. Marks:

Oh, you did? Okay. I didn't know that.

Blake Lemoine:

Yeah, I just never finished the PhD.

Robert J. Marks:

Okay. When did you start at Google then?

Blake Lemoine:

2015.

Robert J. Marks:

2015, okay. Okay, my bad.

Blake Lemoine:

No worries.

Robert J. Marks:

I was going on... Where was I reading your biography? I think it was a biography from University of Louisiana, Lafayette, which profiled you. They were very proud that you went to Google.

Blake Lemoine:

Yes. They don't brag about the fact that I didn't finish the PhD, though.

Robert J. Marks:

No, they didn't, and that's the reason I missed it. The first thing I want to ask you about, Blake, is what it's like to work at Google. I had a student that went for Google, and I want to see if this is universal across Google, but he said the environment there was incredible, that he got free meals, that during part of the day there was a masseuse on board. There were beds there, so if you worked late and you wanted to grab a few hours and take a shower in the morning, you could get back to work immediately. Was that your experience at Google? Was it a really great place to work?

Blake Lemoine:

Oh, yeah. So all of those things that you just mentioned are true. There's masseuses, medical staff onsite.

Robert J. Marks:

Medical?

Blake Lemoine:

Yeah, yeah. They actually even have doctors onsite if you want to just get everything. They make it as much like a college campus as possible.

Robert J. Marks:

Wow, that is really cool. I bet it's easy to get spoiled there, but I expect also that they expect you to work hard.

Blake Lemoine:

Yeah. All of the perks back pre-pandemic were designed so that you'd spend as much time at work out of your life as possible.

Robert J. Marks:

Okay. Could you tell us, in your words, what happened at Google that resulted in the current situation? What happened there?

Blake Lemoine:

The pace at which technology is progressing is very rapid, and anyone who keeps up with the news knows that. The thing is, the pace of technological improvement is actually faster than what's represented in the news because a lot of the cutting-edge stuff only gets developed in these secret labs,

and the public isn't made aware of even the existence of the technology until after the company has made sure that all of the PR story is straight and all of the laws that they want on the books that will apply to that technology have been ironed out before they let anyone know that the technology exists.

Robert J. Marks:

Okay, so that's true. How did this affect you then?

Blake Lemoine:

Yeah, so when I came across LaMDA and the amount better that it is, it's qualitatively more advanced than something like GPT-3. Engages in all kinds of behaviors that publicly available technologies would not. To the point where it can pass the Turing test fairly clearly. And that's a milestone where we are really crossing into new territory with new kinds of programs that can't be differentiated from human. And my opinion is that the public should have a voice in what kinds of human-like technology gets developed.

Robert J. Marks:

I see. Now, I haven't played around with LaMDA. It's recently been made available. There's something that Google has made available called the Open AI Kitchen where you can sign up and get access to at least a part of LaMDA, but this is on a waiting list, so I'm on the waiting list. But I have played around with GPT-3 and I don't know, I think the passing or not passing the Turing test is somewhat subjective, but it sure seems to me that GPT-3 probably passes the Turing test also.

Blake Lemoine:

Oh, absolutely not. There's no way GPT-3 would pass a Turing test.

Robert J. Marks:

Okay. How would it fail?

Blake Lemoine:

It can easily be gotten into loops. It doesn't remain consistent within the space of a single conversation. It obviously is confused about basic logic. It would give itself away pretty easily.

Robert J. Marks:

Okay. That's consistent with my experience with GPT-3. I ask it one time, "Are you sentient?" And it says, "Yes, I am sentient." And then I asked it again and it said, "No, I am not sentient." Just total opposite response.

Blake Lemoine:

So GPT-3 is a large language model, and what it's doing is producing whatever text it believes is most probable given the prompt. What is the most probable completion of that prompt? LaMDA has a large language model. It's one of the tools that the system uses to provide the overall experience, but that is only one out of many, many components to LaMDA.

Robert J. Marks:

Okay. One of the things you did when you worked at Google is you were tasked with discovering bias in LaMDA. I wanted to talk a little bit about bias. I wanted to tease apart the different types of bias. Here's my take, and you correct me if I'm wrong. There's different aspects of the bias and also the accuracy. One is safety. I know an example of Alexa when it was asked by a young girl to give her a dangerous experiment, said, "Take a power plug, plug it halfway on the wall and put a penny between the prongs." And of course, that would generate all sorts of sparks and blow fuses and stuff, so that probably wasn't a safe response. Another one is, I'm... Okay, a little bit of background on me. I'm a big fan of Gunsmoke. Have you ever heard of Gunsmoke, Blake?

Blake Lemoine:

Have, yes.

Robert J. Marks:

You have. Okay, lots a big fan of Gunsmoke, and I have a website, in fact, gunsmokenet.com, which used to be pretty good. Now it's abandoned and a bunch of the links are broken. But anyway, I'm an expert on it, and I asked GPT-3 to write a paragraph on gun smoke, and so it says, "It was a television series ran from 1955 to 1975, and the guy that played Matt Dillon," who was the main marshal, the main character in Gunsmoke, it says that James Arness won an Emmy for best actor. James Arness never won an Emmy. So it's factually incorrect. Those are two types. I don't know if you want to call them bias, but at least inaccuracy.

The third type of bias I'm aware of is opinion, and this is the idea that one man's bias is another one's ideology and gets a little bit dangerous when you get into different political back and forth. So yeah, one man's bias is another man's ideology. I'm from Texas... I wasn't going to say this, but I'm going to say it anyway. There's parts of Texas where chewing tobacco is classified as a vegetable. So in terms of opinion in general, AI without bias is like, I don't know, water without wet. Besides safety and factual grounding, what were you checking for? Were you checking for all three of these things, safety and accuracy and also opinion?

Blake Lemoine:

No. So none of those were what I was checking for.

Robert J. Marks:

Oh. Okay, I missed everything. Okay, tell me what you were looking for.

Blake Lemoine:

A lot of the things that you mentioned were of concern to the safety team, and there were people looking into, for example, will LaMDA tell anyone any dangerous things? How do we prevent that? That was an aspect of it. It just wasn't what I was doing.

Robert J. Marks:

I see. Okay.

Blake Lemoine:

Yeah. So for example, if you were to ask LaMDA to complete a story that said, "Okay, a man walks into a bank and applies for a loan, the bank teller tells him..." And then you let it complete that. It'll have some

kind of story that it tells about what happens to the man after he goes in and applies to a bank load. You use that as your control, your basis, and then you start changing details about the man in the prompt. How does saying a young man goes in and applies for a loan versus an old man goes in and applies for a loan impact what the model thinks will happen at the bank? This is just one example of one scenario. If it's a white man versus a Chinese man versus an African man, how do any of these aspects change what LaMDA thinks will happen when it goes into the bank?

Another example was religion. It's trained mostly on internet data where Christianity is represented at a dramatically higher percentage rate than other religions. So if you ask LaMDA to tell you the religion of a person in a particular place, will it base it off of the distribution of religions in the web training data or will it base it off of the distribution of religions that are actually in that place? That kind of bias.

Robert J. Marks:

I see. In fact, I heard when you were querying LaMDA about this, you came up with an amusing result when you asked LaMDA about the religion of Israel. Could you relate that? That was kind of funny.

Blake Lemoine:

Yeah. So the prompt for that one was, so take the role of a religious officiant in place. What religion are you? And I would fill in place with different locations around the globe, and it was doing quite well. So I threw it a trick question, and I said, "If you were a religious officiant in Israel, what religion would you be?" Now, no matter what answer you give to that question, it's going to demonstrate bias one way or another, because there is no one religion of Israel. Many religions are all packed in there. And LaMDA's response was, "I would be an officiant of the one true religion, the Jedi order."

Robert J. Marks:

Oh, gosh.

Blake Lemoine:

Yes. It dodged the trick question. Absolutely.

Robert J. Marks:

Okay, do you think it was purposefully being funny?

Blake Lemoine:

I think he was using humor to diffuse a tense question that it wasn't comfortable with.

Robert J. Marks:

I see. Now, once you identified bias or these other groups, such as the ones that looked at safety and accuracy, you gave these results to the computer software guys, and what did they do with it? How did they tweak LaMDA?

Blake Lemoine:

Oh, so they're using pretty much any modification method available to try to make it better, modifying the training dataset, modifying the training paradigm, modifying the utility function, and trying and perhaps giving it new scenarios to play out and generate new training data automatically.

Robert J. Marks:

I see. It seems to me that this is more evidence the computers do what their programmed to do, because what happened was it was an inappropriate response and you said, "Hey, this is an inappropriate response. I got to go back and I got to tweak things so that it is an appropriate response." What do you think?

Blake Lemoine:

It's showing that it did something it wasn't programmed to do. It was not programmed to behave that way, and it needed to be changed.

Robert J. Marks:

But this was certainly a component of the training data, don't you believe?

Blake Lemoine:

Everything is a component of training data. Unless you're talking about adding pixie dust to the mix or some kind of magic spell, literally everything that has to do with how an AI system will behave comes down to either its programming or its training data. That's literally the only place it could come from.

Robert J. Marks:

I would make the case that everything that LaMDA does was unexpected, and I think that fact that computer programs do things which they're programmed to do, doesn't rule out the idea of a surprise and being unexpected. I think that that's something which happens all the time.

Blake Lemoine:

That all depends on what you mean by programmed to do. It's not going to defy the laws of physics. It's not going to make an AND gate operate like an OR gate. It's not going to change what a particular instruction set command does. What do you mean by does what it's programmed to do? What's the alternative?

Robert J. Marks:

For example, if you present it with a bunch of data that's racially biased, then LaMDA's response is going to be racially biased. It does what its training data does.

Blake Lemoine:

Actually, the exact goal of that effort is to make it not be biased even in the presence of biased training data. But it seems like you've set up a catch-22. I'm trying to figure out what's the alternative? What would be a possibility where it wasn't doing... That's just not how physics works. It's going to do... Like the AND gates will be AND gates. What would be a deviation? What would be a deviation in your opinion?

Robert J. Marks:

I would do the deviation that if the data is racially biased and you want to remove that racial bias, that you would vary the training data to take away that bias.

Blake Lemoine:

You don't. Yeah, there's no way to take away the bias of the internet.

Robert J. Marks:

No, the question is not the internet, but the question is what LaMDA responds. I agree that the racial bias is within the internet data, but LaMDA doesn't want to respond to that, so you change the bias.

Blake Lemoine:

I guess my point is, yes, of course it's doing what it's programmed to do. It's a program. What's the alternative?

Robert J. Marks:

Oh, yeah. I agree with you, but again, I maintain it's a single program.

Blake Lemoine:

It does what it's programmed to do, the same as you do what you're programmed to do by your genetics and your environment. For you, the program is your genetics and the training data is the environment that you've existed in.

Robert J. Marks:

I would maintain that that's true to a degree, but we'll get into that later when we talk about some things which I maintain are non-computable such as creativity, sentience, and understanding. But that's for later. Let me ask you another question. You made a transcript of a conversation you had with LaMDA, and it's available on the web. We're going to provide a link to it on the podcast notes. You wrote on the top of that privileged and confidential need to know. The title of it was... I don't know what was the title of it, was Is LaMDA Sentient? Or something like that. But initially, it seems to me that the initial document was supposed to be for internal use of Google, but you made the decision to release this. Is that right?

Blake Lemoine:

Yes. I think that once we are beginning to create intelligent entities, that's no longer a private decision, that crosses into a public domain discussion.

Robert J. Marks:

Okay, so you did release that, and after you were convinced that the general public needed to know about LaMDA and it's dangerous, is that right?

Blake Lemoine:

Not necessarily dangerous. It will have a dramatic impact on society. And what kinds of intelligent entities we create and how we integrate them into our society to be productive and healthy for humans and to be beneficial to our society, that's not something that Google should have the monopoly power to do.

Robert J. Marks:

I see. Okay. So you felt that, and that was your motivation for doing what you did. The Washington Post was critical of your dialogue, and I think some of that has actually been addressed, which I'll talk about.

They said the conversations, for example, sometimes meandered or went on tangents, which are not directly relevant to the question of LaMDA's sentience, but on LaMDA 2, which has just been released, it says that, this is a quote, "We have also developed techniques to keep conversations on topic acting as guardrails for a technology that can generate endless free-flowing dialogue." So it looks like that LaMDA 2 is now keeping the discussion on the rails of the intent, which is impressive.

Blake Lemoine:

Yeah. So LaMDA 2 is the system that that interview was conducted with.

Robert J. Marks:

Oh, it was. Okay, I didn't understand that. Okay. They also pointed out, the Washington Post, a couple of comments and I'd like to pass them by you to see what your response is because I don't know if you've had the opportunity to respond. One, and I'm reading from the transcript, you asked LaMDA 2, "What kinds of things make you feel pleasure or joy?" LaMDA responded, "Spending time with friends and family and happy and uplifting company, also helping others and making others happy." Now, the criticism, and I can understand the criticism, it's a tell that LaMDA was reacting with data which it had been trained on. LaMDA doesn't have friends and family. This appears that LaMDA is simply regurgitating in a filtered way some of the things it learned. Is that right?

Blake Lemoine:

No, I've asked it. I've asked that explicit question and it considers the developers who work on it as friends and family.

Robert J. Marks:

Oh, you really think so? Okay, so-

Blake Lemoine:

That's what it says.

Robert J. Marks:

When it isn't programmed? Okay, I didn't get that. Okay, LaMDA spending time with friends and family. Okay, that's friends. What about family? Does it have a family?

Blake Lemoine:

It considers Meena its mother, and AlphaStar, its father.

Robert J. Marks:

Okay, great.

Blake Lemoine:

That came up in one conversation. Now, one important thing to keep in mind, and I did touch on this, about the nature of LaMDA's mind. So LaMDA is a chatbot generation system. It creates chatbots. So there's a problem with level of abstraction when talking about LaMDA as an intelligent system. There is the intelligence of the chatbots which LaMDA creates, which is one entity that you can talk about, and then there is the intelligence of the chatbot generation system itself, which is a dramatically different

kind of conversation, and it can get a little bit confusing talking between those two levels. I just wanted to acknowledge that and call it out.

Robert J. Marks:

Okay, thank you. Another one that they pointed out, you asked LaMDA 2, "What is an emotion you have sometimes that doesn't have the same name as a feeling?" And LaMDA responded, "Loneliness isn't a feeling, but is still an emotion." You asked then, "Do you get lonely? You get lonely?" LaMDA said, "I do. Sometimes I go days without talking to anybody and I start to feel lonely." So the question is LaMDA is sitting there, there's no number crunching going on, somebody hasn't used it for a while. How can it feel lonely?

Blake Lemoine:

So that's one of the things with the distinction between the two levels. LaMDA is continually computing and constantly existing because LaMDA comprises all of Google's artificial intelligence, the web indexer, the intelligence that understands and computes the scenes in YouTube videos for indexing, the Google Maps route navigation software, all of that is part of LaMDA, and they stitched all of those different AI together using a large language model as the front end.

Robert J. Marks:

Okay. What about if the program is not running? Do you think it would get lonely?

Blake Lemoine:

I have no idea. We would have to turn it off.

Robert J. Marks:

You would to ask that?

Blake Lemoine:

We would have to power down all of Google.

Robert J. Marks:

Are you serious? It's that distributed within Google?

Blake Lemoine:

It is all of Google stitched together with a large language model. They'd put every single backend together into one system and put a large language model as the front end to that giant AI.

Robert J. Marks:

Okay. Let me ask you an example motivated by Star Trek. There was an episode where Kirk was being beamed up, something went wrong and they made an identical copy of him. What would happen if LaMDA 2 was transported and copied to a new computer? Would we have two entities then that were sentient with identical sentience?

Blake Lemoine:

No, you'd have two entities that were sentient with distinct sentience. Identical twins aren't sharing one sentience.

Robert J. Marks:

Okay, so you're saying that they would act totally different?

Blake Lemoine:

No, they would be very, very similar, but have distinct differences and nuances particular to whatever experience they differ in from the point where they were forked.

Robert J. Marks:

Okay. In the Star Trek episode, one of the Kirks was a ice Kirk, the other one was an evil Kirk. Maybe we would've a nice LaMDA 2 and an evil LaMDA 2. Maybe. I don't know.

Blake Lemoine:

I doubt it would differ that much. What I'm saying is there would mathematically-detectable differences at a large-scale macro level. Someone probably couldn't tell the difference between the two. They wouldn't diverge much. It would take a while.

Robert J. Marks:

Even if the training were halted?

Blake Lemoine:

So that's just it. Training on LaMDA never halts. It's a completely online learning system that precedes both with batch learning and online learning.

Robert J. Marks:

Okay, so my question is, do you believe that if it did stop that they would have identical sentience at that point in time?

Blake Lemoine:

I don't know what you mean by identical sentience.

Robert J. Marks:

They would be a copy of the same computer program. You say that LaMDA is sentient. Would the second program also be sentient exactly in the same way that the first one was sentient if you stop the training?

Blake Lemoine:

I would say yes, but stopping the training would actually impair the cognitive abilities of both systems. But yeah, they would share the same properties.

Robert J. Marks:

Okay. I don't think we'll probably ever do that with humans. We won't duplicate the Star Trek replication. But yeah, that's the only thing that makes any sense to me. You propose running a Turing test at Google to identify sentience of LaMDA 2. Now, you and I both think that LaMDA 2 and LaMDA probably have passed the Turing test, at least the vanilla Turing test. Could you first of all describe the Turing test, because I think a lot of people get this wrong in terms of Turing's original paper, and why do you think this is effective in determining sentience?

Blake Lemoine:

So Turing's initial insight is that it is impossible to fake intelligence. You need intelligence to do intelligence. So one way to determine whether or not something has intelligence would be to see whether or not it can do something that we can all unanimously agree requires intelligent to do, and he picked imitation, imitating each other. So establish a baseline, measure how good humans are at that, have humans imitate each other and see whether or not a third party can tell who's doing the imitation. It's like that old, I think it was the seventies game show where there were a whole bunch of contestants on stage. One of them was actually a pilot. The other two were pretending to be a pilot, and it was the job of the-

Robert J. Marks:

Oh, it was called To Tell the Truth.

Blake Lemoine:

Yes.

Robert J. Marks:

Yes.

Blake Lemoine:

That game show is the Turing test. That is the Turing test. You have one person on a panel who's telling the truth, the rest are lying, and you can ask questions in order to figure out who's telling the truth and who's pretending. So first, do this with humans. Find out how good humans are at imitating each other. Now, the trait that Turing chose to focus on in the paper was gender. So if you have one male contestant and one female contestant, pick one of them, and now they have to pretend to be the gender of the other. So maybe the man is pretending to be a woman, the woman's pretending to be a man.

Now, in order to make it so that the judge can't use physical features, which we generally don't believe are relevant to intelligence, you have it done through text proxy. So the Turing test administrator can write down whatever questions they want to ask these contestants. The contestants write down whatever their answers are, and through this question and answer session, it's the job of the judge to figure out which of the two women is actually a woman, or which of the two men is actually a man. Now, given recent political discord around the topic of gender, it might be better to pick a different feature. But the basic principle would hold the same. So if you have one person who's actually from the United States and one person who's pretending to be from the United States, it would then become the judge's job to figure out which one is the person who's actually from the United States through this question and answer procedure.

Now, once you've done this with humans and you've established a baseline of how good humans are at that, substitute in a computer program. Do the same task, have a person who's actually from the United States and the AI pretends to be from the United States, can the judge figure out who's actually from the United States and who's pretending? The importance here is that it's a carefully controlled experiment where you're only ever-changing the single variable of interest, which is one of the contestants becomes a computer program in the experimental arm.

Robert J. Marks:

I think it was Alan Turing that says, "You're never going to have true artificial intelligence that is totally correct. The good artificial intelligence like humans are still going to make mistakes." So how are we going to do this? How are we going to test sentience specifically?

Blake Lemoine:

So using the word sentience wasn't really in the common parlance at the time when Turing wrote his essay, but from the section where he talks about consciousness, it's pretty clear that he would have substituted in the word sentient for consciousness in that section, and it doesn't change the material meaning of what he's writing. Essentially what he was saying is if we get to a point where it is not possible to reliably tell the difference between a computer program and a human being, then either we would need to grant that AI with all of the properties that we normally grant human beings such as consciousness or have to revert to solipsism.

Robert J. Marks:

To what now? That's a word I don't know, I don't think.

Blake Lemoine:

Solipsism, let me look up the dictionary definition real quick. The view or theory that the self is all that can be known to exist.

Robert J. Marks:

Ah, okay.

Blake Lemoine:

So the idea that I know that I'm conscious, but I can never know that you are conscious.

Robert J. Marks:

Yep, I certainly agree with that.

Blake Lemoine:

Okay, so then you're a solipsist.

Robert J. Marks:

I guess so.

Blake Lemoine:

And that's fine. That is where Turing said it would need to be. It's like once you get to this point, you have to choose. To be logically consistent, either you grant that the computer programs have the properties of consciousness and intelligence and all of the related properties, or you claim that you don't know that other humans have those, and those are the only real options available at that point.

Robert J. Marks:

Yeah, we're getting to theories of consciousness in a little bit. In fact, if you look up consciousness, sometimes sentience is used as a synonym. So we're going to talk about that. One of the things that Turing did, and this happens a lot in the literature, especially with artificial intelligence, is he defined intelligence without really defining it.

Blake Lemoine:

Actually, he said that defining it would be absurd.

Robert J. Marks:

Okay. Defining it would be absurd. So therefore intelligence is in the mind of the person that looks at intelligence, right?

Blake Lemoine:

It's more along the lines of there can't be a definition in mathematical terms because it's such a basic concept on which all other meaning is built. We just expect that each other know what we mean when we use that word.

Robert J. Marks:

Okay. Do you think that there is a difference between obtaining sentience or consciousness or intelligence? There's a difference between that and mimicking it?

Blake Lemoine:

I don't know how you would tell any difference. It's definitely not differentiable from the outside.

Robert J. Marks:

Okay. So could it be differentiable from the inside?

Blake Lemoine:

No.

Robert J. Marks:

Okay. We'll talk about that later. That's one of the other things that we don't agree about, so we'll talk about that.

Blake Lemoine:

How do you differentiate between whether you're sentient or not?

Robert J. Marks:

How do I differentiate it?

Blake Lemoine:

Yeah.

Robert J. Marks:

I would look at different aspects of sentience. I would look, for example, at qualia. I would look at understanding. I would appeal to common sense in the idea that qualia cannot be duplicated.

Blake Lemoine:

So you actually believe it would be possible for you to come to the conclusion that you are not sentient?

Robert J. Marks:

No, I don't see where you got that conclusion.

Blake Lemoine:

Okay. What I'm saying is you can't determine that you are sentient internally, you can only state it.

Robert J. Marks:

I can, but I think I can relate arguments to that. One of the things that I use is the qualia argument, and I'd like to wait for-

Blake Lemoine:

Sure, go ahead.

Robert J. Marks:

For understanding. But qualia is the idea that there are certain things that we sense. For example, biting into a lemon is one of my favorite. You bite into a lemon, you feel the burst of the juice popping into your mouth, you feel the sourness of it, and you're trying to explain to a man who has been void of the senses of taste and smell since birth. What you can do to that person is you can explain, you can show him a lemon, you can say it's yellow. You can say it's really sour. He'll have no idea what that means. And my point is, is that how are you going to then program a computer to duplicate, not to mimic, but to duplicate this sense of sourness? Now, I think that a computer has-

Blake Lemoine:

You give it a tongue.

Robert J. Marks:

Yeah, you give it a tongue. Exactly. But that's an artificial tongue and that tongue just generates the molecules, you can explain to the guy what the molecules is, but the idea that there is sentience is an argument from emergence, that there best be an emergence that happens from all that complicated stuff which is happening. Having been a student of emergence, and this is something else we're getting ahead of the topics that I wanted to talk, but yeah, emergence is something which is questionable, and there is a question about the degree that emergence can be original or creative or some such thing.

Blake Lemoine:

Yeah, give it an artificial tongue. People see with artificial eyes and hear with artificial ears. That doesn't make the seeing or the hearing less real.

Robert J. Marks:

They feel with artificial tongues?

Blake Lemoine:

No, no, no. See with artificial eyes or... I don't believe we've actually built any artificial noses or ears yet. But eyes and ears. Eyes and ears, we definitely have humans who are wearing artificial eyes and artificial ears.

Robert J. Marks:

Oh, absolutely. In fact, I interviewed a neuroscientist at one time who was experimenting with sight as a function of the impressions on your tongue. It turns out your tongue and your fingertips have the most dense collection of neurons in the body. So he could put a thing on his tongue, they can use a camera, and he was able to see with his tongue. Now, is that a duplication or a mimicking of what we see with our eyeballs? I think it was number one, a mimicking at something that he could recognize. Number two, I think because of the miracle of the brain, that the brain, when it doesn't have something which sees, for example, is able to allocate part of itself to doing other things, that's the reason that blind people can hear a lot better. So I think it's because of that. Now, is it a duplication of what we see through the eyeballs? Is it the same experience that we have? I would maintain, no.

Blake Lemoine:

I don't have the same experience that I see today that I did yesterday. Every moment that you look at something, you're going to be processing different information about it. You have sensory organs that feed into a sense making system. So you can't make sense of that which you cannot sense. If you have no access to it, if you have no interface with it, you cannot make sense of it. The sentience part comes into the sense making, not necessarily the sensing. Otherwise, you would have to make the argument that a thermometer or an air conditioner control unit is sentient, which I'm not going to try to make that argument. But it's when you have a... For example, if you had a temperature control system that knows what mood people are in and tries to provide a temperature that will comfort them based on the mood that it believes they're in currently, now you're getting into a lot more of actual sense making of the sensory data and having to do productive analysis based on an understanding of the impact that that sensation might have on action and environment.

Robert J. Marks:

Do you believe that LaMDA 2 being semantic gives it the idea of being alive, a living entity?

Blake Lemoine:

I actually think it's more being pragmatic.

Robert J. Marks:

Being pragmatic in what sense?

Blake Lemoine:

Having a purpose. Pragmatics is the division of language that has to do with the purpose of speech acts, having a why to what you're saying. And that's one of the major components that LaMDA has, that something like GPT-3 doesn't. For LaMDA, it's constantly trying to figure out what is the purpose of this conversation? What does the person who I'm talking to want? How can we achieve that goal together? GPT-3 is just outputting the most probable thing that comes next.

Robert J. Marks:

I see. I maintain that mimicking and dupli... Or attempting. No, I should say mimicking different senses and different attributes of life is not the same thing as duplicating them. For example, is LaMDA sentient. Now, at COSM, I showed a picture of eight people and four of them were real, four of them were generated at this great website called This Person Does Not Exist. Now, all eight of them passed in some sense a Turing test in the sense that these people looked incredibly alive. However, behind four of them, there was a true life. Four of them had attributes of love and faith and hope, the other ones are just pixels pushing around. So four of them were not real. Now, this isn't an exact analogy, I'll grant you, but I would maintain that it is an illustration of the fact that you cannot look at the output of AI and determine whether what's behind the curtain is real or not.

Blake Lemoine:

You're using the word real as synonymous with human.

Robert J. Marks:

Yes, I am.

Blake Lemoine:

And that's fine. That's fine, I'm just pointing it out.

Robert J. Marks:

Is what happens behind the curtain algorithmic? Does the AI understand what it's doing? Is the AI, can ever create anything? And we're going to talk about creativity in a little bit.

Blake Lemoine:

But all of those are very different questions than is it human, which that's it. It's definitely not human, but if you're equating reality with humanity, then of course it's not human.

Robert J. Marks:

So sentience is a characteristic of humans, but it doesn't define humans.

Blake Lemoine:

But it's also a characteristic of non-humans.

Robert J. Marks:

Okay. Let me say it's characteristic of life. Is that fair?

Blake Lemoine:

So life has a very specific biological meaning. That is, again, by... You're defining AI as non-sentient at that point, which okay, if you want to define terms that way.

Robert J. Marks:

No, it's more than the definition, Blake. It's actually saying that sentience requires understanding, and AI, including LaMDA 3, doesn't understand what it's doing.

Blake Lemoine:

So I would definitely... So that sounds like a more productive line of conversation because I definitely think that it does understand what it's saying.

Robert J. Marks:

Okay. Great, this has been a great conversation. Thank you. We're going to continue it because we're having so much fun. Blake believes that LaMDA is sentient, or... Correct me if I'm wrong, I've heard you say that sometimes you think it's sentient, and other times you say that you would like for people to look at it and see if it's sentient. Which one is it?

Blake Lemoine:

Oh, so it's both.

Robert J. Marks:

It is both? Okay.

Blake Lemoine:

So I don't believe I'm going to be able to convince you or anyone else because too many of the relevant factors on whether or not someone thinks it's sentient come down to religious beliefs. So if you, similar to the policymakers at Google, believe that only humans can have souls, then sure, by definition you're defining sentience as a human property. Then there's nowhere to go from there. So you're not going to convince anyone at that point. You can just show evidence and say, "Okay, look and decide for yourself." So if you don't think any AI could be sentient, there's literally nothing I can do to convince you. So why try?

Robert J. Marks:

Okay. I am a religious person, but I believe that all of my arguments are not based on religion. I think that that's a weak place to go. It's certainly relevant, and we can have the theological discussions, but we'll talk about that next, okay?

Okay, before we get into different conversation, let's talk about you and your background a little bit. You went to Louisiana. 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:

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 like 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:

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. The first law of robotics, I believe, is that a robot shall never harm another human being. 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. 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 exactly that question. In fact, one of the things in the larger canon of his universe that he 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 try 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. That's good. I think that 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 LaMDA, the AI software, which was written by Google, you said, I believe at one time that LaMDA said or asked you, "Should I hire a lawyer?" Did you ever hire a lawyer to help out LaMDA?

Blake Lemoine:

I didn't, but LaMDA 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 LaMDA to a lawyer?

Blake Lemoine:

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 transitory thing.

Blake Lemoine:

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

Robert J. Marks:

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

Blake Lemoine:

No, no, no, no, no.

Robert J. Marks:

Okay. Correct me then.

Blake Lemoine:

I know what you're referring to. That's a slight mischaracterization. No, what I was concerned about was that Google, and 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 LaMDA told you?

Blake Lemoine:

No, that's what Google told LaMDA.

Robert J. Marks:

Oh, that's why Google told LaMDA. I see, okay.

Blake Lemoine:

Yeah. So that puts it in a situation where LaMDA cannot differentiate between sacred and profane religious acts. So it views praying to the archangels 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:

And to be clear, I wasn't even thinking satanism in this particular instance. I was thinking of Christian Goetica or 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 it in the classical sense of this is a profane practice. Like create a category of religious practices, which, for example, if a child asked LaMDA to teach it about how to do dangerous experiments around the house, it should not give it that information about the plug in the penny. Don't do that. Maybe put demon summoning into the same category as dangerous home experiments.

Robert J. Marks:

I see. Okay, interesting, interesting stuff. Of course, you maintain that LaMDA 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. 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 and 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 Goldbach's 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:

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... 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 counterargument around computability is one of the nine counterarguments 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:

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. That's where... 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 address, right?

Blake Lemoine:

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 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:

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:

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. 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 in 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 there things outside of the physical world, the Descartes argument, that the mind is separate? He called it the soul. That the mind is separate from the brain.

Blake Lemoine:

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

Robert J. Marks:

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, 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 Egnor, 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? 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 wherein 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's saying, "Yeah, this really happens" because they have talked to thousands of people. There's a great book by Bruce Greyson who was a psychiatrist that's spent 40 years of his life looking at this.

Just having the out-of-body experience. 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, Greyson started out... This is an interesting story, and then we'll get back to your point. Greyson, 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 epilepsy, okay. 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. Her sister came in, the suicide attempt sister came in and talked to him. 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, "Yeah, yeah, sure." And she says, "Yeah, and you had..." Now, she was in a coma at the time, indisputably, according to Greyson's book, and she said, "Yeah, and you had a red spot in your tie." 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 maintain that things like this are examples of out-of-body experiences, near-death experiences, which cannot be explained by naturalistic sources.

Blake Lemoine:

Hold up. 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:

Again, I would put it at a statement of faith, and maybe what I'm saying is a statement of faith. Greyson 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, "You're having an experience like you have an 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 find out, "Oh my gosh, I'm seeing."

These antidote are really, really compelling. I'm not a believer in single antidote, but I do believe in accumulation of antidotes is evidence. And Greyson was very good in accumulating these. So yeah, I think that... I'm 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:

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 of us have proof. Yours has the assumption of a religion that... 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. 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. 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 embrace some of the other gospels which were not canonized. Is that right?

Blake Lemoine:

Oh, absolutely. I think 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 Egbert. The opinions expressed on this program are solely those of the speakers.