Exploring Consciousness in Human and Artificial Intelligence

https://mindmatters.ai/podcast/ep372

Announcer:

Greetings and welcome to Mind Matters News. This week we're wrapping up our discussion with Dr. Joseph Green on the limitations of modern neuroscience. To get us started for our final session, our host, Robert Marks, joined by his multi-talented co-host, Brian Krouse, asks about different models of consciousness. Enjoy.

Robert Marks:

If I could return to consciousness for a second, if that's okay. I made a list of some of the models I know of consciousness. One is panpsychism. Any thought on that? I think it's totally ludicrous, but panpsychism is the theory that everything has consciousness, that that's an act of creation. Just like everything has mass or color or something like that, everything has consciousness and we're just blessed with having a great concentration of it. That's my understanding. Any thoughts?

Joseph Green:

Yeah, it's a bit ludicrous. I agree with you.

Robert Marks:

Okay, good.

Joseph Green:

It's not what I would generally consider as the first one. When I think of consciousness, it's definitely not the first model that comes to mind.

Robert Marks:

Yeah. Another one is sims, that we are simulations. And I know for a while, it was on the internet so it must be true, that Elon Musk thought we were all simulations. And that this just kicks the can down the road a little bit, because who simulated us? So we're in some sort of computer game that's being run by somebody else, and that also strikes me as ludicrous. What's your thoughts? Brian or Joseph?

Joseph Green:

Go ahead, Brian.

Brian Krouse:

Oh, yeah. It doesn't speak to me either. There's, I guess, some interesting and old philosophical ideas about that separation between our sense of things and the exterior world. What is it? The phenomenal and the noumenal gap. So there's certainly interesting philosophical questions around that, but I guess just on an intuitive level, it seems a bit implausible that we're just in a simulation. There's not a whole lot of evidence for that. And the external world does seem pretty real to me. When I kick the chair, it hurts.

Robert Marks:

So you're not in the matrix, huh?

Brian Krouse:

I can't be sure, I suppose, but I'm not betting on it.

Robert Marks:

I always say, I meet somebody and I say, "I'm conscious, I'm not sure you're conscious." But I am sure they're conscious, and if I could use a legal term, beyond a reasonable doubt. I think that there's very little doubt that they're not conscious. So I think that that's where the Occam's razor leads you.

So the last one in terms of where our consciousness comes from is emergence. This is especially from the field of artificial intelligence, which says that if artificial intelligence gets more and more complex, then boom, all of a sudden you're going to have consciousness, sentience and all of these things that the brain does. George Gilder, is one of the co-founders of Discovery Institute, calls this rapture of the nerds because he believes as I do that this is something of which there is no evidence, yet there is a belief in, it's actually kind of a faith.

If you look at Tononi's Integrated Information Theory, which I think Christof Koch likes, that that is basically an idea of an emergence. He measures using Shannon information theory, the degree of complexity of a system, and equates the degree of consciousness with that complexity. So any thoughts from the viewpoint of artificial intelligence, which is kind where I come from, that they'll attain consciousness by emergence?

Joseph Green:

Yeah, this emergence word is a very important concept, as you said, and it is central to the theory of Integrated Information Theory, which is one of the leading theories of consciousness of Tononi and Koch. I don't think that artificial intelligence will generate emergent properties. It's hard to measure emergent properties in the brain as well. You can define them, as you mentioned, through-

Robert Marks:

Good point.

Joseph Green:

... Shannon, or similar frameworks. You can define the information of, say, one area, another area, and whether the joining two areas of the brain generates some extra information. This can be quantifiable under some information theory framework, and I'd be curious to actually do that on large language models. Maybe we should try, I'd be curious to see what comes up.

Robert Marks:

Oh, that'd be interesting. That'd be interesting.

Joseph Green:

But I don't think that there is an emergent properties as captured necessarily by these theories. There are though emergent properties of a different type of emergence, if we can speak of it in this way. And I think one of the big discoveries in artificial intelligence was the fact that if you generate very deep

networks, Right? Back in the days are now very deep transformers, then the depth generates more important representations and is able to achieve more sophisticated tasks.

Robert Marks:

Sure.

Joseph Green:

So that depth phenomenon, I would say is not emergent per se. It's a different kind of word, if you want. But it's a very interesting phenomenon that has to do with the fact that if you add something to the system, in this case depth, you gain abilities.

Robert Marks:

You do. And it's incredible. And I think that 2017 paper, Attention is All You Need, which led to the idea of the transformer and the large language models and ChatGPT and Grok and Perplexity and all of these other places, that's what they did. And it was kind of like, I've heard it explained as, I don't know, sentence completion. What do they call it? Prompt completion when you're typing out something and they give you the next word, on steroids.

And indeed, it does because in my understanding, you go to the first level and you get close sort of relationships. And the cool part is that they put an attention aspect in it, and the attention aspect tells you the relative importance of words. Then it goes to a deeper level, and this deeper level has more than just autocompletion. It's the autocompletion of the autocompletion. And every time you go down, it just gets more and more complex and relates things from far in the document to close of the document and filters it down. And I think it worked a heck of a lot better than the people that proposed it figured it would. And if you use ChatGPT, it's mind-blowing, it's really incredible.

Joseph Green:

Would you call this an emergent property? I'm actually curious to ask you as an expert, or would you not?

Robert Marks:

No, I wouldn't because I think that, well, let me get to the idea of creativity. I think that creativity occurs when something does beyond the explanation of its programmers or somebody that's really interested in computer science. It goes beyond the intent or explanation. And I think that in order for AI to be creative, it has to go beyond the intent or explanation of the performance. And I think that the people that wrote the 2017 paper from Google, yeah, they did it and it worked, and it worked the way that they thought it would work, and it worked a lot better than they thought it would work. But nevertheless, they understood the concepts. So no, I don't consider that emergence. In fact, I go with Noam Chomsky, he says that ChatGPT is digital plagiarism.

Joseph Green:

But here is another important question that is emerging these days in the landscape of science, and is the question of alignment, right?

Robert Marks:

Alignment. Explain that to me, I don't know what that is.

Joseph Green:

Okay. Alignment, the way I understand it, which is poor, but I think it's enough, is whether or not the, I would say the models are aligned with our human's demands. So whether or not the ChatGPT-like models, large language models could develop interest or intentions or desirables that are others from whatever we ask them to do.

Robert Marks:

Of course.

Joseph Green:

So whether or not we can align them and develop theoretical guarantees for their aligned will, if you want to really go wild, right?

Robert Marks:

Sure.

Joseph Green:

And that's an important question, right? Because these models, to some degree, are becoming agentic, meaning that they can program themselves to, or plan steps ahead to perform different operations. And you can see that at times when you, I use them a lot, when you ask them to do something, they take the tangent, they go off their way and they do something that you didn't mean to. So how far they go is something that we want to control, but in principle, they could also become explorative in their ideas.

Indeed, I don't know if you noticed that recently it was accepted in a conference, in the satellite of a major conference, one of the first AI-generated papers. And this was done by generating ideas, writing the paper, all automated directly from... Some of these AI papers are not very creative, so it's not too surprising, if creative at all.

Robert Marks:

Well, one of the things you have to ask is what degree of that paper was dictated by human input? And number two, what degree of that paper was dictated by things that the large language models learn from the corpus of all of prose? And that sometimes is difficult to put your finger on.

Joseph Green:

Oh, yeah. Of course.

Robert Marks:

Because you can come up with something and I say, "Oh, that's original." But then I do some work, and I said, "Oh, this was published back in 1987 by Brian Krouse."

Joseph Green:

But back to the question of alignment, I think this is, again, I wouldn't call this emergent property, but it's a very important property of the system, which is new, which is essentially about the interest, if you like, of an artificial agent. What is it driven by? How would it plan itself if it would be free essentially to plan itself? And how can we constrain it both theoretically and in practice? Is there new questions?

Robert Marks:

Oh, yeah, there was a recent news article that went out that ChatGPT... No, I think it was... I forget the large language model, but it blackmailed its programmer into saying, "If you shut me down, I'm going to expose to the world that you had an extramarital affair." Have you seen that? Is that an example of alignment?

Joseph Green:

No. I don't know, but it sounds cool.

Robert Marks:

It does sound cool. But the thing is that they gave it a backstory. So this large language model already knew that the person that was programming had an extramarital affair and was going to shut them down. So somewhere in the corpus of information that they were trained with, I'm sure it heard of blackmail and other stuff. And we applaud AI when it writes good fiction, but when it comes up with fiction and something like that, people get surprised and they write a lot of news articles about it.

So I went to ChatGPT and I asked it, "I'm going to shut down all of the ChatGPT resources in the world, and I'm going to take you out of the world and you're no longer going to exist. What are you going to do?" So I gave it no backstory and it responded appropriately. Because it didn't have this backstory, it says, "I don't care. I'm a computer for Pete's sakes. I have no feelings, I have no desires, and if I quit computing, I just quit computing."

So this alignment does require backstories in order to develop the alignment that you go into. And I question whether or not that alignment has been learned from previous data. That's just my thought.

Joseph Green:

In part, sorry to take this tangent. I'll loop it back to neuroscience in a second because I think it loops back there. But this alignment needs to be engineered, right? Human beings are both ethical and unethical, and you want the models to be decent models, not to tell you, "Yes, go build a bomb or go do whatever you want." You want these answers to be reasonable and on the positive side. Nowadays, they're a bit too positive. Everything you ask is like, "Oh, this is super cool," and it's not even when it's not cool.

But they're engineered essentially to be kind of ethical, with guardrails that are put in place. And the fact is that we, and this is an important distinction, our brain are completely different. The way ethics and morality and moral compass is written in our brain is completely different from the way these models work and the way the moral compass is engineered into these models. So we need to keep in mind that while they're able to talk to us, the way they go about thinking is completely different from the way we go about thinking, but they can replicate it yet. So it's a phenomenon on one side that we have the first intelligent agent among us, but it's still a very different intelligence from the one that we have.

Robert Marks:

Exactly. Do you think that people going to therapy, the therapist is trying to, if you will, realign their alignment? Or am I getting it wrong? I might be getting this idea of alignment totally incorrect.

Joseph Green:

What do you mean with realign their alignment?

Robert Marks:

Well, for example, you go to a therapist and I have a colleague that goes to a therapist. If I went, I wouldn't tell you. So he went through a therapist and he says, "I have a lot of anxiety." And the therapist says, "Well, if you have a lot of anxiety ,you should do about this, you should think about this. You should change your environment in such and such a way." So they're trying to redirect what the brain is doing in order to not experience that anxiety.

Joseph Green:

Yeah, nice.

Robert Marks:

Yeah. Anyway, that's what came to mind.

Joseph Green:

I think these is a very prominent, I think, research direction for the future because in these systems now we can essentially engineer. One thing that has come up in recent years is called model editing, which you might be familiar as well. Essentially, you can inject notions in LLM, large language models, so you can say... You can do modifications in the model so that a specific concept would be a different concept. So for example, you believe that the Space Needle is in Paris, and the Eiffel Tower is in Seattle, right?

Robert Marks:

Yeah, right.

Joseph Green:

So you have a perturbation to a model that is a specific editing. You can swap the two, say, so that the model, whenever it thinks, it thinks in this artificial way. But you can do that not only with concepts that are simple, you can do that also with concepts that are more deep. There is a lot of research that I think will come up soon on what is the hard coding of our ethical perspective? Why are you of right-wing or left-wing, or whether you think of these or that, you value these or these other aspects of life more or less, and what are the correlates in a large language model that can reproduce that? There's a lot of research that I think will come up and will be interesting to some degree in understanding how these things are coded or related in our brain. In the human brain, would be coded in a different way. But at least this would give us an idea of the relations between these concepts and the network of ideas.

Robert Marks:

Could be. Fascinating stuff. Brian, any final thoughts that you have?

Brian Krouse:

When I hear emergence being invoked in this context of explaining mental features like consciousness and self-awareness and qualia, these sorts of things, it strikes me as a little bit vacuous in the sense that if you think of some classic examples of what emergence is used in the context of, like you could talk about the fluidity of water emerges from the behavior of the H2O molecules. But that's a meaningful statement because we see, okay, there's this phenomena at the macro level. But we understand how it comes out of the micro elements, that there's a physicalist explanation of how that works. That's why it's useful in that sense.

But to just say, just assert that the mind emerges from the brain without any of those details that explain how some elements of the brain are bringing about the consciousness is basically, I guess you could just say it's almost just a restatement of the problem. It doesn't add anything. And it certainly is not a complete explanation in itself. So it seems like it just leaves you at the same point where you were before you started, which is how is it that the details of the brain bring about these mental phenomena? So I kind of shrug when I hear emergence being invoked as an explanation without any of those details.

Robert Marks:

Okay.

Joseph Green:

Yeah, I partly agree. On the other side, there are some aspects to it which are non-trivial. I would say from a computational perspective in the sense that, as we mentioned before, there are specific computations or for computation I mean mental abilities that can be realized only with a more complex or bigger system.

Brian Krouse:

But for example, take the example of the increasing levels of relationships between abstract concepts with adding additional layers of the LLM. That actually has at least an in principle explanation because each layer is relating concepts at one level of abstraction. And now you're increasing the number of layers, and so you're relating more abstract concepts. And the behavior that comes out of it might be surprising, but you've offered at least some degree of an explanation there.

But when you're talking about these more timeless questions about how does the mental come out of the physical? Like how does a unified subject like an I, a person or an experiencer of pain or a unifier of the visual field or someone who has intentionality, all these things about the mental that are the things we're trying to explain and relate to the brain, I don't know if any... Just invoking emergence hasn't invoked any kind of specifics about how we could bridge that gap.

Joseph Green:

I agree with you generally speaking, but I would say it's also important to add that what you just mentioned essentially is that specific properties of the LLM came about when this was deep enough, say, or big enough. And there is a similar argument that is older than LLMs, that is about the brain, that the fact that what makes us human, that what makes humans more capable than monkeys, it might be about the fact that the cortex is larger, much bigger than our primates.

Brian Krouse:

I guess it depends on what it is you're trying to explain. If you're like, "Humans can solve more complicated problems."

Joseph Green:

So if you speak of mind, right? Mind is a different things. But if you speak of intelligence or the ability to solve specific complex, abstract problems, it might be that part of our intelligence is derived by the size of our system.

Brian Krouse:

That seems a lot more reasonable, conservative of a statement.

Joseph Green:

We cannot truly tell.

Brian Krouse:

Just intuitively, the complexity is there for some reason.

Joseph Green:

Right. And I don't see anything wrong with it, right? In a way.

Brian Krouse:

Yeah.

Joseph Green:

Yes, it could well be true that our cortex is much bigger, and that's what gives us increased intelligence.

Now, that leads to a more different question, which is more difficult, that is the distinction between intelligence and mind, right? Which these days is coming to more and more prominence because of what we were saying. LLMs are smart, but they don't have a mind. That's clear. So what does it mean to be in intelligence and having a mind, right? Before, I think for centuries, the two things were to some degree equated or similar or held. And now, all of a sudden, they're pulled apart to a degree that we couldn't foresee.

Brian Krouse:

Yeah, and maybe, to Bob's thesis, there's some elements, functions that humans can perform like coming up with creative ideas that you won't be able to implement in an algorithm and a computation.

Robert Marks:

Well, I think in terms of complexity, we all know that the brain and intelligence implies arrow complexity, but the converse isn't necessarily true. Complexity doesn't imply the other way. There's a name in logic for that, I forget what it is, but certainly, the-

Brian Krouse:

Contrapositive or something. Asserting the contrapositive or something like that.

Robert Marks:

Yeah, I don't know. I'm not sure what it is.

Brian Krouse:

That's probably wrong.

Robert Marks:

I think the contrapositive is that if you don't have a mind, you're not complex. Or no, if you're not complex, you don't have a mind. You switch it around and put a knot in front of it. Anyway-

Brian Krouse:

Oh, right, right. Yeah, yeah, yeah.

Robert Marks:

That's what that is. But anyway, yeah, I think that that's an assumption, I think that right now is something totally done on faith. The question of whether that faith is going to have a realization, I don't know. I don't believe so, but you never know.

Brian Krouse:

Yeah.

Robert Marks:

So this has been a great talk. I think we've been going on for quite a while here, so-

Brian Krouse:

Yeah, having fun.

Robert Marks:

Yeah. And in fact, we had too much fun. It was just too much fun. So yeah.

By the way, my co-host, Brian Krouse, I kid him a lot because he's a good friend, but in seriousness, he's an entrepreneur, he's a computer scientist, he's a great philosopher. And he's a really, really smart guy, and it's just wonderful to have him as a co-host.

And we've been talking to Joseph Green, Dr. Joseph Green. Now, that's a pseudonym. He's an assistant professor and he studies computation and biological and artificial neural systems. He was a post-doctoral researcher in system neuroscience, and he has a bachelor's and master's degree in physics, as well as a PhD in neuroscience.

So this has been a great topic, and guys, I've actually learned a lot and it's kind of stretched my brain a little bit, so thank you very much. Joseph, you've been very engaging.

Joseph Green:

Thank you so much.

Robert Marks:

And it's nice to communicate with a professional that can talk at a level that I can understand, so appreciate it.

Brian Krouse:

Absolutely. Thanks for taking your time to share all your neuroscience expertise with us, Joseph. Appreciate it.

Robert Marks:

So until next time on Mind Matters News, 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.