TOMASO POGGIO: I'm Tomaso Poggio. Of course, did not change. What we had to change, unfortunately, for the sad events happening in Israel right now is that Amnon Shashua was in the panel, cannot attend because they are in a state of war right there. And Pietro Perona has been so gracious to accept to jump in at the last moment, so he will replace Amnon.

So there will be three of us real and three virtual. And I don't want to spend too much time in-- at this point, I think we have a fantastic panel, I think will be very interesting to ask all of the panelists about their opinion on various topics, including some that were discussed already.

The questions I want to address are, to start, two. And I asked our panelists to speak for about five minutes about both of them, and then we'll continue the discussion among them. The first question is about whether it's a good idea to push more work on theory.

Theory comparing large language models or other deep learning models and human intelligence, and comparing large language models among them because we are at this unique point in the history of mankind in which you can argue that we have more than one intelligence around. More than human intelligence.

And so I think it would be very interesting to see whether there are or not common principles of intelligence common to this different forms of intelligence.

And the second question is one that has been the core throughout this workshop and is essentially whether neuroscience can help AI and AI can help neuroscience.

So to give you a small idea of the kind of fundamental principles I'm thinking of about, deep learning and human learning is a question related to puzzles existing at this point, the main one being, as you know, there is a curse of dimensionality.

If you want to compute a function, do a computation that requires more than a few inputs, you incur in this problem that you potentially need an exponential number of parameters to do a good job in representing the kind of computation that you want, the kind of function that you want.

But apparently, neural networks are not really subject to this curse, the question is why and whether this gives us some insights about them and potentially other forms of intelligence.

OK. And there are other questions related to that having to do with the difference between classical and quantum computers, but we'll discuss that later.

So the first-- I would like to ask Geoff. I think Geoff is, apart from me, maybe the oldest in the group. I met Geoff for the first time-- I think was '79, Geoff, in La Jolla. Do you remember?

GEOFFREY HINTON: Yeah, I remember it.

TOMASO POGGIO: All right. So we are old enough.

[LAUGHTER]

And so that's a good argument for having you start this discussion.

GEOFFREY HINTON: OK. So, taking into account age is the only thing I've got going for me--

[LAUGHTER]

I am going to ignore the question about theory. I don't do theory and my math isn't good enough, so I'll just focus on the other two questions, the role of neuroscience and AI and the role of AI in neuroscience.

I think neuroscience is clearly had a huge impact on AI. Just the idea that big networks of simple processes that learn by changing the connection strengths, that's-- the idea that they can do really complicated things that's not initially very plausible, and it's because of neuroscience, that people explored neural nets at all. So that's the main influence.

There's lots of little things like ideas like dropout came from thinking about neurons and how neurons occasionally don't spike when they should. Ideas like ReLUs came from both Sam Roiz and Hugh Wilson telling me that sigmoids really weren't very good models of neurons. ReLU was a much better model of a neuron. Even though logistic sigmoids have very nice Bayesian interpretation, but they don't work as well as ReLUs.

One other influence that I think hasn't happened yet and needs to happen, and it hasn't happened for hardware reasons, is fast weights. So Ilya will be very familiar with the idea that you can actually implement fast weights, but it doesn't buy you much because you can't do your matrix-matrix multiplies anymore because every case has different fast weights, so you can't share the weight matrix between different cases.

I think maybe people will eventually start using fast weights when they have different hardware where maybe weights are conductances. So that's something that's yet to come. But I had an epiphany recently. This spring, I was thinking about analog computation and how to make things more energy-efficient.

And I suddenly realized that maybe these digital intelligences we've got that use backprop where you can make many copies of the same model that behave identically are maybe actually better than what we've got. They're still a bit smaller than what we've got, but not that much smaller. They can pack a lot more knowledge into a lot fewer weights.

And my current belief is they're probably just a better form of intelligence that couldn't evolve because it's so energy-intensive. It needed us to create it. But my current belief is that-- so this switch is the other question, is the role AI for neuroscience, but new developments in AI may not be telling us much about neuroscience.

It may be we've got most of the information inspiration we could from the brain, and now the new developments don't necessarily tell us much about the brain. That's a radically new thought for me. For 50 years I thought that if you made your AI a bit more like the brain, it will work a bit better, and I no longer really believe that.

One other thing I think AI told us is that these big language models-- or big multimodal models, especially, have undermined the kind of argument that the brain is more statistically efficient than artificial neural nets. And I think what they've shown is that people previously were comparing what an MIT undergraduate can learn from not much data with what a tabula rasa neural net can learn from not much data.

And if you take a large language model and you can pick the large language model against the MIT undergraduate on some new task, there's not nearly so much difference in statistical efficiency. This is the few-shot learning.

So I think AI has taught us something about why we're statistically efficient, and I think Bayes said it rather a long time ago, that we have a very, very rich prior, but there's no reason why digital intelligences can't have a similar enriched prior. And I think I've used up my five minutes.

[APPLAUSE]

TOMASO POGGIO: So, Pietro. Pietro Perona is a professor at Caltech. He's a renowned for his pioneering contribution to computer vision and computational neuroscience. He has played a key role in advancing object recognition, texture analysis, and vision science research. This is, by the way, written by ChatGPT.

[LAUGHTER]

Slightly bombastic, but OK. And yet shaped the future of vision technology also engaging with the next generation of researchers such as Fei-Fei Li. And I think you are a VP or something at Amazon in your spare time, correct?

PIETRO PERONA: I-- not VP . My title is Amazon Fellow, and I spend half-time there--

TOMASO POGGIO: OK.

PIETRO PERONA: --at the moment. Mostly working on responsible AI at Amazon. So hi, Geoff. Hi, Demis. Hi, Ilya. OK, so I don't have prepared comments because I was asked at the last moment, but I can say a few things.

So something I noticed is there is a debate going on on embodied intelligence versus intelligence that you can derive from purely reading text. And so I'm in the camp that thinks that at least part of the things we know how to do have to come from having a body, and in general, intelligence will be different if you're embodied or not.

Right now, we have been overindexing, for a very good reasons, on intelligence we can get out of analyzing batches of data that is available on the internet. But in time, we will want to have intelligent machines out there that act in the world. And so we know about Marc Raibert's robots. We've seen them. We know about cars that have to drive on the freeway and don't yet have a good theory of mind about the drivers, and so they don't interact well with the rest of the world.

Now if we want to go there, then some of our perspective on how to do intelligence will change. Right now, we are poisoned by correlations, and we don't know how to build machines that can understand causation. And to understand causation, fundamentally you've got to allow the machines to carry out experiments. And so why so?

Well, because if you're a non-embodied agent, well, prediction is all you need to do, and correlations are very good for prediction, but if you need to change the world, and so you need to do intervention, then you understanc the causes of things, and then causation is fundamental being able to reason about that.

So we haven't yet seen artificial systems that carry out experiments and design them at the same level even close to what humans can do. I've seen machines built by my students that can beat any naturalist at recognizing plants and animals. You can carry in your pocket iNaturalist and recognize about 100,000 species of plants and animals.

But no machine comes even close to a lousy biologist in thinking about how to carry out an experiment, interpret the results, et cetera. So that's terra incognita. It's very interesting to me. I think we should think about getting there.

So then people wonder about superhuman intelligence, and of course, machines have unfair advantages. They have access to much more information at their fingertips. They can communicate through just batches of parameters what they learned to another machine, which we cannot do, et cetera, et cetera. So they can use more sensors, more able-bodied, if you want. So it's clear to me that in some domains, at least the machines will do better than we do.

And now the third question that Tommy was asking was about theory. And things-- the facts on the ground are changing so fast. We have a moving target, and so of course we want to develop theory, but that takes a long time. Sometimes we've seen shreds. I think that Geoff was mentioning, for example, dropout.

So in my mind, the main most successful piece of theory for deep networks is how to avoid overfitting and how to do regularization and so we understand much better how to do it now. So I would say that's a good piece of theory that has come out. And so it will be in the future a mix of this moving target of what people can do. We have to remember that it's about 50,000 very smart people around the world trying things out and seeing what comes out. It cannot beat that by pushing equations, there's just this massive exploration.

And so theoreticians have to pick the most juicy targets and work there and see if they can keep up. And so we see both theory and experiment, and depending on how things go, the one or the other will have an advantage and we don't know which one is going to be. So I think I will stop here.

TOMASO POGGIO: Thank you.

[APPLAUSE]

So David Siegel. David is Co-Founder and Co-Chairman of Two Sigma Investment, a large hedge fund. And I think we met in what is now the Moderna Building, and it was the AI Lab. And you were a roboticist. And in the meantime, you were working in a field of finance that I think is probably the hardest for machine learning.

So to be very interested to hear your perspective not only from the point of view of research, but also business, communities going forward in the context of this AI revolution. David.

DAVID SIEGEL: Thank you, Tommy. I am-- I'll touch on that at the end, but let's start for a moment by everyone relax, we're almost at the end of a wonderful 10-year celebration. Close your eyes for a moment. Go back to your childhood when you were very young. Imagine you're with your parents in a field looking up at the stars. Just try to recreate the thoughts that you might have had back then. OK. Now hold that thought and I'll get back to that in a moment.

When I think about my own journey and why the study of intelligence is important, initially it started, really, by an interest in trying to get computers to do human-like things. And that's what brought me to the AI Lab. And I wanted to essentially solve practical problems.

And today, the field of AI has advanced substantially, and it is able to demonstrate an uncanny ability to solve various classes of problems. It's quite remarkable. We can debate how much thinking it's doing, but you can't debate how the results can be applied to businesses like my own. It's infusing various kinds of technologies on the internet. That's all quite good.

But more recently, though, I began to-- and in part, joining with Tommy here when he was conceiving CBMM, began to reflect more on, really, the notion of intelligence, and what does it mean and why is it important to think about it? And when you go back to dreaming about the stars, long ago for some of us, you were wondering, I believe, given everyone here in the room, what's going on up there?

You were not just looking at why are they-- how are they moving? You weren't thinking about predicting orbits. You were thinking big-picture. This is really a remarkable thing. You're looking at what made this be? How does it work? Why is it here? And all of this is really a problem that mankind has studied for almost forever.

So much effort goes into trying to understand our universe, not because we're going to be building some kind of practical device out of it. An understanding of the universe isn't going to make a better search engine. We just want to-- because we want to know why we're here, and this is just part of our existence.

Intelligence is the same thing. To understand our mind is a key, key aspect of understanding our own existence. And if we don't understand what our mind is and what intelligence is, we're never going to really know who we are. And to me, this is a basic research project. It's not motivated by commercial gains.

Understanding intelligence probably will help advance commercial AI applications, which is wonderful, but that's not why I'm interested in the problem. No more than the cosmos will not lead probably-- maybe it'll lead to a commercial application, but I doubt most researchers in that area care about that.

I wanted to reflect a little bit on the actual question of neuroscience and AI. Well, if you're framing the problem the way I'm trying to frame it, the answer is they have to work together because the problem is to understand-- I think the grand problem is to understand a theory of intelligence and how the brain is able to compute in a way that provides intelligence and extend it to consciousness that makes it too hard to even comprehend. We can stick with intelligence for the time being.

This theory of intelligence, it's not going to come, in my mind, from an experiment-- purely from just, oh, it looks like it's intelligent. We need to have a deeper understanding to be able to essentially describe what's going on just the same way as to-- we could use some model to predict the motion of the stars, but that's not good enough for understanding what's going on.

That's not-- when you were dreaming as a kid, you wanted to understand, well-- I mean, you didn't know that, maybe, but how is gravity working, what's holding the thing together, how did-- you needed the theory.

So I believe that what we should be-- everyone doesn't have to work on this, but CBMM, its focus has been largely on what I'm describing. It's not meant to be, let's make commercial AI better. I think that's good. There are plenty of people, you have them on the screen here, who work on that problem and are doing a great job. I think what we're doing here is much more about what I'm describing.

And I think that the-- frankly, the amount of money that is invested in this kind of research is unbelievably small compared to studying other grand challenges just to help us understand who we are.

So the neuroscience part and the AI part I think are the best way to develop combining them together to understand this theory of intelligence, the theory of the mind. I'm not sure it will lead to better AI. Perhaps we've gotten all the-- as was suggested a few moments ago, perhaps most of the benefits have already been achieved to be inspired-- inspiring AI, but to me, that doesn't really matter.

And so I'll conclude just simply by saying, I think, really, this is all about understanding who we are, and to do that, we need a theory of intelligence.

[APPLAUSE]

TOMASO POGGIO: Could not agree more. Very good. So now Demis, Demis is the Co-Founder, CEO of DeepMind Technologies, and more than that now. And as you all know, it has made pioneering achievements in the virtual world of games with AlphaGo, AlphaZero, and more recently, in the world of science with AlphaFold. He received the Lasker Prize recently. Congratulations, Demis.

[APPLAUSE]

You probably did not receive it yet officially?

DEMIS HASSABIS: Yeah, no, a couple of weeks ago. It was a great ceremony.

TOMASO POGGIO: Was it? OK. Yep. And before that, he was a game designer, but essentially a neuroscientist with a PhD in neuroscience in university/college, which I find it's a very interesting fact that you are heading arguably one of the best AI companies. And another one, Google. And you are a neuroscientist, not a computer scientist. So Demis, up to you.

DEMIS HASSABIS: Thanks, Tommy. So I just want to-- actually, Geoff said a lot of the things I was planning to say, and I agree a lot with Geoff's opening remarks. So I think neuroscience has made a huge contribution to AI. It's more subtle than just pointing at, oh, which algorithm? I think it's just the soup of ideas.

Not only in deep learning, but also in reinforcement learning, which I think has been important. And other things, too, even smaller ideas like memory replay, episodic memory, things like that.

The seed of those ideas has come from our understanding of neuroscience. And even if they end up being implemented in a way that's quite different in silico.

So I think we shouldn't underestimate that, and I think the neuroscience community should feel very proud of that. But we are moving into a new era now, I think, where it's now becoming very engineering-heavy what we do in the cutting edge of AI, and the systems are increasingly diverging from maybe how the brain works.

So I've came to that realization maybe probably four or five years ago as the scaling laws seem to start holding with these what have become now the frontier large language models. And in fact, they're not just language anymore, but multimodal.

Now having said that, I still think there are missing things with the current systems, but if we were-- and we've had these discussions over many years over the last decade, and if we were to be talking five years ago or something, I think would have made arguments about how are we going to-- how are these systems going to learn abstract concepts or form abstract conceptual knowledge, including eventually symbolic knowledge like language? Maybe we will be discussing about grounding in the real world through embodied intelligence or simulations. And it just turns out, there is this other way to do it.

But probably because-- it's a little bit-- I regard a bit like the Industrial Revolution where there was all these amazing new ideas about energy and power and so on, but it was fueled by the fact that there were dead dinosaurs in the ground, and coal and oil and just lying in the ground. Imagine how much harder Industrial Revolution would have been without that. We would have had to jump to a nuclear or solar somehow in one go.

And I think that's what's happened with intelligence research and AI research, and the equivalent of that oil is just the internet, this massive human-curated artifact that we've been building over the last 20, 30 years, the whole of humanity has been building.

And of course, we can draw on that. And there's just a lot more information there, I think, it turns out than any of us could can comprehend, really. And that's what these massively-scaled AI systems are able to draw on.

So I think that allows us to-- along with some human feedback, by the way, that reinforcement learning human feedback, which think some grounding seeps through that, in effect, because we're obviously grounded agents, we're interacting with these AI systems. Perhaps they're ungrounded, but then they get grounded feedback. So effectively, there's some grounding, I think, seeping into their knowledge and their behavior through that approach.

Now-- so there's still things I think that are missing. I think we need we're not good at planning. We need to fix factuality. I also think there's room for like memory and episodic memory. So I think there's still a lot of room for inspiration to come from perhaps neuroscience ideas, but perhaps the peak of that has passed now.

On the other hand, what I do think we should be taking from neuroscience is analysis techniques. So I think that's what I would-- and I've talked to Tommy about this-- this is what I pitched CBMM on, is that we're really lacking in our understanding of these systems. It's incredibly hard to do because AI is an engineering science. You have to build the artifact first, which is already very hard, before you can then decompose it scientifically and understand it. And that's a moving target, obviously.

So it's a very empirical science, and I think we need an empirical approach, as well as theory, but empirical-- but largely empirical approach to trying to understand what these systems are doing. And I think that neuroscience techniques and neuroscientists can bring to bear their analysis skills to this.

So the kind of work I have in mind is things like Chris Olah's work at Anthropic, I think some of the best examples of this, of analyzing what these systems are doing and the representations and the architectures and so on, but I just think we need like 100x more research on that.

So that's got to be the goal. I think the other thing that places like CBMM can do is the leading labs, I think-- and including ours, would be willing to give early access and access to these very large models for the purpose of analysis and red teaming. And think critically, we also need benchmarks, to benchmark capabilities. And that obviously has safety implications as well as performance implications.

So for example, it would be good to know if these systems are capable of deception, but what does that mean? How can we operationalize that in a way that can be tested, rigorously tested and perhaps passed in some sense?

So there is a huge amount of work here to be done, I think urgent, urgent work to be done that is getting done by some of the leading labs, but there's not enough of it. And I actually think we better if independent academic institutes were to do this and join that effort.

So perhaps I'll end with obviously, as these systems get incredibly powerful, and probably very, very soon, I think it's an urgent need for us to understand these better. I mean, there may be other possibilitiesd for example, like these systems explaining themselves in addition to us analyzing the representations and things.

So I think-- I've got a lot of optimism that we can do this, but it needs a lot more people, a lot more great researchers joining that effort. Thanks.

[APPLAUSE]

TOMASO POGGIO: Ilya Sutskever. You are the youngest kid on the block. The last one to speak. Did great things with Geoffrey first, with ImageNet. And then he was more recently a Co-Founder and is Chief Scientific Officer or something like this at OpenAI. And so responsible for ChatGPT and GPT-4. Ilya.

ILYA SUTSKEVER: Thank you for the introduction. Lots of extremely good points were made, so i'll be concise. So there were three questions being posed. What is the role of theory? How can neuroscience help AI? How can AI help neuroscience? So I'll start with that very quickly.

Theory, theory can mean different things to different people. It is unlikely, due to the great-- due to the very high complexity of neural networks, that we'll be able to have very precise theory at the level of-- that we can make an incredibly good predictions like we do with physics.

At the same time, theory is obviously useful. And you could see even today that once you give up on the desire to have extremely precise theory, then suddenly a lot of ideas around scaling of parameters, scaling of activation, activations, their normalization, theory of optimization, suddenly all comes together, and is obviously extremely useful for AI today. I am completely certain that more theory like this will continue to be useful, so that's the first point.

Second, what can neuroscience give to AI? Indeed, it was already mentioned that very important giant ideas have come from neuroscience to AI. For example, the idea of a neuron. The idea of a distributed representation. Many of them were mentioned before by Demis, by Geoff right now. And it is possible that there is maybe one or two or three more such big ideas which we can borrow from neuroscience.

But it takes great skill and incredible taste to borrow ideas from neuroscience successfully. The brain is immensely complex, and neuroscience produced an incredible an incredibly giant set of facts about the brain, about neurons, about their spikes, and about their ion channels, maybe about the large-scale organization of the brain.

And it is not at all obvious which of those ideas are incidental and if you should not worry about them, versus whether there is one particular idea in the brain that we perhaps can use as inspiration for our research.

So I think it is possible that we will find in some small number of years that something that we've discovered has an analog in the brain or perhaps that the inspiration will go from the brain to our AI systems, but it needs to be done carefully. And I would not say, oh yeah, just go look at the brain and copy it. So that's the second point.

AI helping neuroscience. One very interesting thing, and it would be very, very amazing if it turned out to be true, is that there is this mounting evidence that the representations learned by artificial neural networks and that the representations learned by the brain, both in vision and in language processing, are showing more similarities than perhaps one would expect.

I don't think this is something that was immediately obvious in advance. So maybe we will find that indeed, by studying these amazing neural networks that everyone is producing right now, it will be possible to learn more about how the human brain works. That seems quite likely to me. I think its fruitful.

I also want to echo-- I want to also echo and support Demis' point around some specific useful things that can be done in academia and in this center, one of which would be the evaluation, understanding what these models can do, how good are they really?

It's very confusing. Sometimes you have these models, they show amazing flashes of brilliance on the one side. On the other side, they have these very strange, quiet non-human failure modes. Getting more insight into what's going on there, and even better, trying to gain an understanding of where things will be-- any insight at all about where things will be a year from now or two years from now would be quite helpful, would be a very important contribution to be done outside of the big companies. I'll stop here.

[APPLAUSE]

TOMASO POGGIO: So maybe we can sit. Let's try this arrangement with the virtual ones. I wanted to first to show a couple of slides just to introduce the discussion. Yeah, because our panel-- some people in our panel literally did not hear this yesterday. But this was one reason I spoke about why theory is good. I was making the example of Volta discovering the battery and the fact that that discovery brought really immediately a revolution even if nobody understood what electricity was.

In 20 years after Volta invented a pila-- "pila" means "pile of things," which are discs of zinc, cork, and copper, this thing here. And 20 years after that, there were electrical motors, electric generators, electrochemistry was done, telegraph lines.

So it was a big revolution, but it's only much later that we had an understanding a theory of electromagnetism with Maxwell. And of course, this contributed even more to powering the revolution of electricity. It's 1800, it's not very long ago. 200 years or so.

So that's why theory, among many other reasons having to do with, for instance, what David mentioned before, and you want to understand ourselves, having theories means understanding what's going on. We would like to have a theory. And does not need to be the same level of precision as physics or Maxwell equations. They may be just some fundamental principles.

And then on the second point that was-- what Demis and also Ilya referred to about what a research program could look like would be a kind of empirical study of different forms of intelligence. So large language models, different ones. Possibly comparison with human intelligence. And this would be at the level of cognition and at the level of what is inside the box.

And one of the goal would be to look at systematic differences and common properties, common behaviors, common-- if there is any, leading ideally to some fundamental principles common to these intelligent systems.

There may be none. I personally believe there would be. And this would be maybe not an understanding in the precise sense of classical mechanics, but something equally useful and important.

So let's start from here. He wants to have a first take on this? Geoff, what about more on the theory? You are-- I guess you are on a negative sign or neutral.

GEOFFREY HINTON: No, I'm not against theory. I think theory is great. It's just I don't do it. I'm not very good at math and I'm much prefer doing more practical things. There's something I forgot to mention when I was talking about AI for neuroscience. I think it's made one huge contribution in neuroscience in the last few years, which I think is it's given us a much better understanding of the nature of language in the brain.

So there's this crazy guy at MIT called Chomsky who has been claiming it's all innate. We know now that doesn't have to be the case. And Chomsky's whole view of language is kind of crazy when you look back on it because language is about conveying meaning. It's about conveying stuff. And Chomsky ignored that aspect of it.

It's as if you wanted to understand what a car is, and for all of us, understanding a car would mean-- a large part of it would be understanding how the engine works that makes it go, but you can imagine someone saying, no, no. The thing about cars is, to understand why it is you get three-wheel cars and four-wheel cars, but you never get a five-wheel car. And that's what we need to understand about cars.

And that seems to me, like Chomsky's theory of language, he wanted to understand why certain syntactic constructions aren't possible. And as far as I can see, he did everything he could to avoid the basic issue of, how does language mean? And I think these large language models have put an end to that. Not in Chomsky's mind, but in more or less everybody else's mind.

DEMIS HASSABIS: Yeah, just to follow up on that, and also just mention about AI for neuroscience as well. For what it's worth, Geoff, I've always thought Chomsky was completely wrong from my undergraduate days. Just-- and I think sent natural language processing down the wrong route for a long time. But maybe that's for another day, that discussion.

But AI for neuroscience, I forgot to add that part. I think maybe that's what should happen now, is, we have all these amazing AI techniques-- and I know many people that are doing this, let's apply it across the board to analyzing, decoding brain states, all sorts of things, things we used to do 10, 15 years ago, but obviously we have much, much better AI tools now.

I feel like the-- I mean, I've gone a little bit away from neuroscience in the last few years, but my feeling of the field is-- and maybe people will disagree in the room there, is that we need to start asking better questions on the neuroscience side.

I just don't feel I've seen any results or that have been really major leaps in like learning theory or representations, or maybe I have just haven't seen the papers. But I like work from people like Tim Behrens and stuff like that in Oxford and UCL, but I just haven't seen a huge flourishing of that that I felt was there maybe 20 years ago when fMRI came in as a new tool.

And maybe there's a chance to do that now if you think of AI, the AI systems we're building as new tools, new analysis tools from a neuroscience point of view. It feels like we should be doing different types of experimental neuroscience, perhaps. I mean, that's just a question for the neuroscientists in the room.

So yeah, I feel like-- but then I just want to emphasize again the evaluations and benchmarks point in addition to what you listed, Tommy, on your slide. It's really important that some-- we do that as a field. Like create the right benchmarks, which does require theory as well. Theory of what is it we're creating the benchmarks to theory of emerging capabilities. I don't think there's any theory of these emerging properties, where they come from, how these systems produce those emergent properties.

If we had better theories about that, then I think we could build better benchmarks, and then we would have a better handle of when they might appear.

TOMASO POGGIO: I think my first bullet was about benchmarking or--

DEMIS HASSABIS: OK, maybe I'm--

TOMASO POGGIO: Looking at common and different features, aspect behaviors of machines and humans and-- between machines, but yes. Anybody-- any neuroscientists in the audience want to answer to Demis' challenge? Jim.

AUDIENCE: Yeah. So I think some of us have drank that Kool-Aid a while ago that we're using these things as our best predictors of what's going on and then deriving experiments from those. That creates new phenomena that we post as benchmarks. What's unclear is how to use those phenomena to then turn the crank again to build a deeper understanding in that.

And some of us have also been on the benchmark bang for a number of years, too, which also is not typical in the field of neuroscience and cognitive science. So it's been hard-going, but it's great to hear you guys support both those ideas. And those things take dollars and money, and also a change in mindset I think of how experiments are done and why they're done, not to necessarily create immediate understanding, but to fuel those pumps.

But that next turn of the crank, beyond the experiment and the predictions and whether-- it doesn't-- the experimental effort doesn't amend itself towards the incentive structure of the field where a lab is supposed to produce a result and then a deep understanding of it.

So doing science at scale through platforms like that I think is where we need to head, and I hear that in everything you guys are saying. And I think that's a great opportunity for us to treat the AI generators as hypothesis-builders and us to shape them into which one is most like the brain, let's track that way. And that requires those benchmarking platforms and those experimental things to run together to make that reality.

So I'm just channeling back what you said. Just trying to say, I'm-- I already-- I'm on board with that Kool-Aid, and I hope more people could be, and if you could help us in some ways, that would be awesome. Thanks.

DAVID SIEGEL: I'll comment on a slightly different direction from what Demis was saying related, though. And earlier today, someone made the point that studying an AI system, a large language transformer architecture that is displaying intelligence, is a lot easier than studying the brain because you can poke around at it, you can get whatever you want and analyze it without much effort.

It seems to me that the brain is much more complex than a transformer. And if we can't really figure out what's going on in the transformer architectures at a level that is allowing us to say, oh, now we really understand what it's doing, it'll be almost impossible to do it for the brain, which is going to be much more complicated.

So I see a benefit of essentially shifting some of the neuroscience work to these architectures that are displaying intelligence.

PIETRO PERONA: So if I could add my perspective. I think we may be, at the moment, overindexing on language as a field of interest. And we should not forget that humans are just one species out of very many who exhibit intelligence. And a fundamental principle of science is to study a phenomenon in the simplest possible embodiment so that one can get to the bottom of it more quickly and start from the simpler version to understand the principles.

And so neuroscience has been doing this. And so we do C. elegans about order of 100 neurons. We have fruit fly, order of 100,000. Mouse, order of 100 million. Human, 100 billion.

And we should not forget, all of these different trade-offs between power, consumption, performance, adaptability, and view the problem of intelligence in the context of many different species, if not all of them. So that's something that neuroscience can contribute to keep our attention on different forms of intelligence.

GEOFFREY HINTON: It's interesting that different people have very different views about that spectrum. So I was once talking to Steve Pinker, and I asked him, suppose we understood exactly how a rat worked-- we understood everything we could possibly want to understand about a rat. Would we be more or less than halfway towards understanding human intelligence?

And I think most biologists would say you'd be most of the way there. Steven Pinker said, oh no, you'd be much less than halfway to understanding human intelligence.

PIETRO PERONA: Yeah. Well, that's a matter of taste. So we don't know actually, right? Once we are-- once we are there. But I think it's-- you agree, Geoff, that it's worth keeping our attention on all of these different forms and not obsess about humans and language?

GEOFFREY HINTON: Absolutely, yeah.

TOMASO POGGIO: Mark.

AUDIENCE: Yeah, I've been wondering if psychophysics should be brought into the discussion in a more direct way. I think everybody's talking about neuroscience as though it-- neuroscience as though it has to do specifically with the physics of neurons and interconnections and that stuff.

But the behavior of humans and other animals as reflected in psychophysical experiments, and I think there's an opportunity to get the benchmarks that way, that could be applied both to engineered intelligence systems as well as biological systems that would really provide an opportunity to do comparisons without having to go all the way to the neuron or to whatever the description of the computation is.

TOMASO POGGIO: I-- can I have the slide for a second? Yeah, I think the first point here, the first bullet is psychophysics or cognitive science or measure of behavior. Benchmarking including. The second one is more like recordings. And I think the first one is the more important in this cases.

AUDIENCE: But I wonder what the panel, when they referred to neuroscience, what they meant.

TOMASO POGGIO: Let's ask them. The virtual panel.

DEMIS HASSABIS: Psychophysics. I totally agree, by the way, that psychophysics is actually exactly what we need. And we even-- maybe we were slightly premature with this, but about maybe even almost a decade ago, we had a thing called Psych Lab, which is like a virtual test lab for AI systems.

And I think it's precisely that. The behavioral testing under very rigorous controls is probably something we need to push a lot harder on as opposed to just the neural recording equivalent. So I totally agree that psychophysics should be a huge part of it.

GEOFFREY HINTON: So one question for clarification. If you look at something like AlexNet, it bases most of its decision-- it relies a lot on texture. And if you look at these new AI generative models, they rely much less on texture for doing classification. Is that the kind of thing you call psychophysics? Yeah. In that case, yeah, I think we need a lot more of that.

DEMIS HASSABIS: But it would be also other things, Geoff, like testing like memory situations and setups-- like practical little experiments. Actually, we used to model them on rat experiments originally, but you might have to update them now because our systems are too sophisticated.

ILYA SUTSKEVER: I think one of the nice things about the fact that these very powerful models exist is that some of the ideas that are being discussed here about using psychophysics as an inspiration, like we don't need to make a discussion, we can just go and try it, and very quickly already get some interesting results to discuss.

And by we, I don't just mean people in the big labs. I mean, there are powerful open source models now. There's model access-- like big labs provide model access for researchers. You could find out very quickly.

PIETRO PERONA: So it may be interesting to hear from OpenAI and DeepMind, how do you allocate your resources? So you clearly have obvious commercial industrial goals, and yet you will not be successful unless you do good work somehow. So how do you look at your resources between just technology, theory, and neuroscience? How do you view that within your companies?

DAVID SIEGEL: Maybe one way to think of it-- I don't know if you would agree, but in industry in general, there tends to be more hill-climbing where you bet on a particular approach, and then you keep improving it. In academia, there's more jumping around because your academia is constantly being pollinated with new thinkers, with new ideas.

And there's no infrastructure, really, for hill-climbing anyway. Not scaling things up. So this is, in my mind, a natural division of ideation.

ILYA SUTSKEVER: So the question implied that there is tension between the needs of the product and the needs of research. And there is some sense in which it is true. There is a different sense in which it is a lot less true. And I want to explain the sense in which it is less true.

It is pretty obvious that there is a fair bit of competition between the different companies and how well their AI is doing, which means that if you become a little bit too shortsighted, then next year or in two years, your AI will not be doing as well. And so that creates a lot of desire and simple commercial incentive to continue to improve the AI.

Improve doesn't also mean to make it more capable. It also means to make the near-term AIs more safer, as well as to do work on making our longer-term AIs, especially ones that will be smarter than people, and those AIs will be built, too, by the way, superintelligent AIs, to make them safe, aligned, and generally positively predisposed to humanity.

But how do you do this? How do you work on this long-term research? And there is no easy answer. There's basically two answers. You can hire a lot of great researchers and give them freedom. This is one approach that can be done. Another approach is if you have correct top-down ideas, you're confident, then you could reduce your search space and make progress this way.

And that's basically-- it's like, how's the philosophy? How are we thinking about how things should be as opposed to merely how they are right now? I think all these things that factor in together to continue to make progress.

PIETRO PERONA: So, OK, may I ask one more question? I mean, agitate another question. So I want to go back to something that Demis was saying before, which is we need to invest more on testing and understanding how to test, which resonates very much with me.

I was involved in the field of vision with defining problems through benchmarks, and that worked for a while. Now I find that when I think about these large vision and language models, it's becoming more and more complicated to define what is the task and then what is the benchmark that we should use to measure it.

And so I feel like many of us at the moment don't have a good compass to decide whether what is going on is better or worse than before. And when you think about the life of a scientist in a company or in a university trying to decide if they're doing better or not, they often rely on fairly simple-minded benchmarks that you find in somebody's paper somewhere and you don't even know if they mean anything.

And looking at the neuroscience aspect, I think we also have a little bit of that in the sense that we are interested in understanding how the brain works, but many of us in the labs end up with very stereotyped preparations where the animal has to perform a task that is even unclear whether it has any ecological meaning for the animal.

And the animal overlearns it over months, and then we study what the neurons do. And it's very unclear what it does in our perspective of the ecological value of intelligence.

And so it feels like in both fields, we need to rethink what is intelligence good for? What is behavior? What are animals or automata trying to achieve? And how to measure, in some ways, the fitness of-- ecological fitness of these creatures.

So it feels like it's a very rich set of questions that Demis brought up, and I'm wondering if-- I don't know. So we know what Demis thinks, but I'm wondering if Geoff and Ilya have thoughts on that, whether they agree with what I say, that it's difficult to measure performance.

ILYA SUTSKEVER: Yeah, so there is no doubt whatsoever that measuring performance is extremely difficult. I want to give some examples. You may have heard claims, those of you who have been in AI-- so let's say in the mid-2010s, claims around superhuman performance and vision being achieved.

Like at some point, some researchers have achieved superhuman performance on vision, on the ImageNet data set. Well, then-- but we were obviously not superhuman at this task. How can that be?

Well, but it wasn't really too big of a deal because these neural nets, it was just an academic project-- a research project that some very motivated and passionate individuals were working on, it didn't matter. Now we have much more sophisticated neural nets, they are being used widely, and it is difficult to understand their performance.

If you give it-- if you take, let's say, any one of these-- any one of the large neural networks-- large language model chatbots that you can find online and you ask them to solve a hard math problem and they solve it, is it because they reasoned and understood? Or is it because they saw something somewhat similar or perhaps moderately similar in the training set?

And training sets are quite large. And this creates confusion. You may see people posting really excitedly cool examples of behavior. They go viral. Then other people try to do similar things and they fail. It's not to say that our neural nets don't work, they obviously do. But it does say that measurement is genuinely not straightforward, and this is an area where I think there is room for very meaningful, conceptual, and empirical contributions.

GEOFFREY HINTON: One little comment I have. I used to do experiments on GPT-4 before it could look at the web when I was fairly confident that everything it knew happened before January of 2023-- or was it '22? And so you could log off, log on again, and then ask it a slightly varied version of the question and get a different response. And I don't think you're going to be able to do that anymore.

As soon as anybody talks about doing an experiment-- GPT-4 will be able to see discussion on the web about that experiment is my guess, it's going to be very, very hard to do any kind of systematic experiments.

DEMIS HASSABIS: So Pietro, maybe my thought on this is, look, if it was easy, it'd have been done already. So it's definitely not easy because all the people on this call-- and I've been thinking about this for-- we've all been thinking about 15, 20 years and we have a lot of thousands of researchers and whatever. It's very hard.

But my point is, this is the grand-- this is my pitch for to Tommy of what I believe-- this is what I do if I was at MIT and CBMM right now.

Don't worry about building this compute race of building the largest models. You can act-- I think most of the leading labs would give you access to the models for analysis and safety work-in and so on. So assume you have that. So you don't need to join that race.

What we really need, and what I think you're hearing from everybody from the leading labs, is this is-- what-- and this involves theory, involves neuroscience, involves psychophysics, involves practical experimentation is, how are we going to wrestle these questions of emerging properties, right benchmarks, testing these new types of intelligence as Ilya says.

We had-- we've all seen this. Like in AlphaGo, you have a system, it's better than the world champion, but if you get it out of distribution, you can make it do weird things even today. I mean, we could fix that, but there's no point just for Go. But it's just-- these are lumpy intelligences that seem they can have big delta holes in what they know because of the way that they're trained and the way that human intelligence can't have because of the way that humans learn.

So it may even require a whole new theories or meta-theories of learning that we don't have today. So I think it's extremely rich space for the next five, 10 years, which probably plays into the strengths of what MIT and CBMM can do. And think it's extremely badly needed and urgently needed, and is probably complementary to what the leading labs are good at.

I mean, we try to do a bit of that at Google DeepMind. We do have some neuroscientists that's pretty unusual already for AI labs, but there isn't enough-- they're not attracted to do that kind of work in those kind of places. So I think this is a big opportunity, and it's, I think, desperately needed if one thinks about deploying these systems and the safety of these systems and grappling with that over the next decade as we get closer to AGI or human-level intelligence.

So I think there couldn't be a clearer, in my view, mission or clarion call. And in addition, that will help, of course, understand the human mind in many different ways we've discussed today as well. Anyway, I'll stop there.

DAVID SIEGEL: I'll add just quickly--

TOMASO POGGIO: David.

DAVID SIEGEL: I agree with everything that has been said. Benchmarking is extremely difficult, in part because even if it's a well-defined problem-- so certain problems have a right answer and you can benchmark, but many, many-- most problems don't have a right answer. They have-- it depends on context, for example, or it depends on-- if you're talking about this early, your philosophical framework is overlaid onto--

So there really needs to be some thinking on what we even mean by benchmarking when you're dealing with fuzzy outputs and-- I don't really know. And I would add to that that benchmarking is really important because if you decide on your benchmarking function, then people will tune essentially their models to maximize-- they should-- their performance on that.

And so if you've got the benchmarking function wrong, you could end up building things that aren't that are good at the wrong things, so to speak.

TOMASO POGGIO: Very good. There are many other questions that I want to ask, but before that, let's open to the audience the possibility to ask questions. So let's see. There was-- somebody was the first one. Well, I'll go randomly. So Jean-Jacques.

AUDIENCE: Yes. So this has been alluded to before, but maybe let me ask this a little more directly. So we've been all raised with the idea of evolution, and particularly the fact that the brain evolved from the refinement of sensory motor control. There's the vignette of the sea squirt which is this little animal which swims and then fixates itself on the rock, and at this point, swallows its own brain because it doesn't need it-- doesn't need to do motion anymore.

So here we are at a strange place now with LLMs where we have language before motion. In a sense, we have conversational agents and so on, but we know we're nowhere near to having a robot plumber, let's say.

And it's true that LLMs have, of course, absorbed, as Demis would say, all the knowledge of the internet and so on, but for instance, we are-- I am, certainly, incapable of describing the details of what I'm doing, what I'm manipulating something. What am I sensing? What am I-- so this is not something which is easily describable. So are we missing something by having this kind of language before motion behavior?

GEOFFREY HINTON: I think it's not entirely true. I mean, maybe you could comment on the Rubik's cube.

ILYA SUTSKEVER: Like I think the analogy that the data that exists as a fossil fuel is great. And that's why this fossil fuel played an important role in making the AI that we have today. And at least for now, at least up until this point, in the past, robots were expensive and there wasn't any robotic data. It was expensive, no one could run-- train big neural net and run them on robots. So the recipe that's been working today did not apply to robots.

Changing very quickly. If you look at some of the advancements in robotics that various labs are producing, it's looking pretty good. I mean, there's really cool work from Google, DeepMind, from-- recently from the Toyota TTIC, Toyota Technology Institute, they've been training really cool neural nets that control robots, like flip pancakes and stuff like this. And now it's becoming possible-- people now believe that it's possible, whereas before they didn't.

So indeed, it is true that one could argue that perhaps that AI we have aren't quite as complete and certainly not feature-complete as a robot, but if you look at the advanced progress in robotics, I think in a few years, things will look very, very different then.

DEMIS HASSABIS: Yeah. I agree with that, Ilya. I mean, we just released our RTX generalist robot system, and it's still far from anything that's generally capable of robotics, but I wouldn't hold out that this is different from-- I think language was already very hard and that was doable. And we have other ways to get around it. I think very realistic simulations, physics simulations.

And also, just gathering more data from large robot farms, arm farms. I think there's ways to get around this, and then generalizing from general models and into the robotics domain. I don't think it's going to be that hard in the next few years, so I wouldn't overindex on embodied being different in my opinion.

AUDIENCE: For those of us who sit-in the intersectionality of industry and academic work, when we think about how to apply theories into practice to drive innovation, and hence, lives, blah, blah, blah, in the real world, here are some questions I want to-- one question I want to ask is, we all know the importance of framing a question, a problem. And we talked about I think some of the people talks about the holes in assumptions, what happened in blind spots.

So here's the thing. If you can go back 20 years from-- ago and you look at this AI field, what are potential blind fields-- potential blind spots back then that you may not perceived of that translated to today?

I think for David, for example, when you started Two Sigma, what were things in your journey when starting this company with using machine technology in a very, very imperfect real world-- data imperfection, system dynamics, conflicting motivations between different parties, what are some of those blind spots that may help us as perhaps teaching data and inform and infer how we may look at AI and humans in the next 20 years?

DAVID SIEGEL: Just one quick answer to that. If you're essentially learning off of the wrong data or incorrect data, you're going to obviously get undesirable output. And so if you don't have a theory and you're just essentially being entirely data-driven, you better be very careful about the data. And so that, I think, is a general lesson to keep in mind when things become very empirical.

AUDIENCE: Yeah, I just wanted to ask the panelists from industry how we could push the envelope on really actually bringing large-scale psychophysically-controlled experimentation and alignment research to the ground here in a place like MIT?

So I think it would be no small exaggeration to say that at least in the last five years, what I've observed in this community is the biggest contribution to neuroscience, slash, AI research has been Facebook's PyTorch models. Those tend to be the models that we use in our research. We attempt to try and draw experiments across the open source-available versions of those models.

But I think it would be quite beneficial to be able to do both large-scale experimental grid searches over, for example, psychophysical parameters in an experiment, but also smaller-scale model psychophysical searches where we're able to train much smaller models than something like GPT, but would still need to do it at large scale to train many different versions to test many different hypotheses.

It's like Hubel and Wiesel were doing a grid search, but on a screen with a little pen and a piece of paper. If we could do that at the scale that industry is training models, we might actually be able to pursue alignment research at a much higher clip.

So I'm wondering what the doors to that might be, how you're thinking about that, what resources you might be willing to devote to that, and how we could get that conversation rolling on how to do experimentation at scale with robust controls and checks from academia, but with the resources of industry.

DEMIS HASSABIS: Well, I mean, I can just give a short answer to that. Look, I think the leading labs are willing to-- and we're talking to government about this as well, provide access to the models.

So we should just think about that as a starting point. And some of those models can't-- there's a whole question about open sourcing, which is a little bit out of scope today, which is to do with-- obviously we're big proponents of open science and we published many, many things in the past that's underpinned a lot of the advances you see today.

But as these systems become more and more powerful, we have to answer the questions, like obvious questions, in my opinion, like bad actor use case, bad actors getting hold of powerful technology and then repurposing them for bad ends. And bad actors could be individuals or nation-states.

So one has to have an answer to these questions while still obviously keeping the open science flowing. So it's not easy. There's another thing that's incredibly difficult, otherwise it would have been solved already.

ILYA SUTSKEVER: Yeah, I'll briefly comment on this as well. It is the case that OpenAI, and I believe many other AI labs, provide access to their models for academic research, and that's really the answer. Live models are expensive, but you can still do a lot of things with them. Certainly much easier to run, let's say, psychophysics-inspired experiments on models compared to human beings or rats or something like this.

AUDIENCE: My name is Manolis Kellis, I'm a professor at MIT in AI and computer science, and my research is on genomics, computation biology, and also a lot of molecular neuroscience. So the molecular underpinnings of human disease.

So a lot of us in this meeting have been wondering a lot about the diversity of human neurons and how vulnerability is associated with schizophrenia and neurodegeneration, Alzheimer's, et cetera, are, in fact, pointing to very specific neuronal mechanisms, very specific subclasses of neurons.

And there's a big debate, and today, we talked a lot about that, about how much does it really matter that we have different types of neurons in the brain? How much should we try to understand the role of this extraordinary diversity of dozens of different types of excitatory inhibitory neurons, the role of glial cells, et cetera?

Is AI, at this point, in your view, completely disconnected from that? Do you think that there's just an evolutionary, weird byproduct of the way that we had to get to where we are today that led to this extraordinary complexity of the brain? And if we were to start over with just a giant cortex or something, then we would have been just as intelligent?

And related to that, I mean, we're talking a lot about embodied intelligence about the role of emotions, the role of having the convergence of multiple sensory inputs. The ability to memorize through these engrams, if you wish.

So I'm just curious. In your view, is human intelligence just only useful for understanding the human? Or could there be some true paradigm-shifting capabilities that could emerge from understanding how this bag of noodles has achieved what it takes-- a factory of energy to achieve in terms of cognition?

And, of course, the bag of noodles may be left behind, but at what expense in terms of energy? So I'm curious about that back-and-forth question that I think was part of Tommy's premise earlier on.

GEOFFREY HINTON: My guess is that the brain's been highly optimized over a long evolutionary period. So it's got all these different kinds of neurons because it helps to have all these different kinds of neurons, but that it could do pretty well with far fewer types.

Obviously, it needs several types. So things like layer normalization in the AI models were inspired by inhibition in the brain. So there's a little bit of neural diversity in these AI models that comes from that.

But my guess is a sort of Crick's view, which is that evolution is a tinkerer and it's been tinkering for a long time, and it's come up with all these little tricks that are instantiated in different kinds of neurons, but you probably don't need all of those to get an intelligent system.

ILYA SUTSKEVER: I have one quick comment on this, which is, if we take a trained neural network, perhaps a large trained open source model, we may already discover lots of interesting neuron types. In fact, that is very likely to be.

AUDIENCE: So the panel is phenomenal. So I would like to ask you, what is your opinion about AI-enabled Scientific Revolution? Is AI everything? Is something AI cannot be done? So combine AI with the science. Yes.

GEOFFREY HINTON: Demis has already shown it can be done.

DEMIS HASSABIS: Yeah. Well, look, that's been my goal and passion from the start. This is why I've worked my whole life on AI, is to-- exciting moment now where we can apply it to helping us understand the world and the universe around us. So I think AlphaFold is my calling card for that of what we may be able to do.

And I hope when we look back on it in 10 years' time, it will just be the beginning of a new era of AI-enabled biology or AI-enabled science. I think right now, the way I think about this is, looking at all the systems we've built, if you can boil it down to quite simply, really, I think for situations where you've got massive combinatorial search space, and often, there's a lot of things that can be couched like that. Maybe material design, chemistry, lots of things in biology.

And then there's a solution-- there's a-- but there's a solution to that, whether that's the correct protein fold-out of all the possible ways a protein could fold, for example. You've got to find-- you need a model, first of all, of the underlying space so that you can, therefore, search that intractable space in a tractable way to find that needle-in-a-haystack answer.

And AlphaGo is-- basically, that's what AlphaGo is, but in obviously in the game of Go, because it would be impossible to just do the search on its own. You need a model-- some kind of reasonable model-- it doesn't even have to be that good-- of Go and the dynamics of Go and the motifs in Go. And it's the same with AlphaFold.

So I think right now, there are many problems in science that I think, if you couch it like that, could be solved already with existing systems. Let alone the next systems that come along that might be able to generate new hypotheses and things like that themselves. I don't think we're at that stage yet. We have to put in the hypothesis and frame the question and give it the data and build the model and so on. So that's very much a tool right now for human experts to use, which is what we do.

And it's very general. We're applying it to not just biology, but chemistry, to fusion, plasma-containing fusion, and also mathematics and theorem-proving. So there's actually-- I think when you start thinking of it in that way, there are a lot of problems in science that could fit that type of setup.

DAVID SIEGEL: Appropriate problem for what we're talking about today. I've always believed that AI will advance rapidly enough that it will actually help us perhaps solve the problem of understanding the brain, and we should really be applying this full-circle.

TOMASO POGGIO: Yeah, let me take a quick poll among all of the panelists. One question that was underlying what-- underneath what you said and what was discussed many times is, how original or creative are the latest large language models?

Of course, we know that, for instance, AlphaGo did some pretty creative moves when it won its match in South Korea. So, that's possible. But to be very concrete, do you think the existing models or some-- the next GPT-4, say GPT-5 or so, will be able to state a new non-trivial mathematical conjecture? I'm not saying proving it, I'm saying stating it. We think it's possible within the next five years.

ILYA SUTSKEVER: Are you sure that the current model cannot do it?

[LAUGHTER]

TOMASO POGGIO: I'm not sure. Absolutely. Do you know whether it can?

ILYA SUTSKEVER: I mean--

GEOFFREY HINTON: Let me give you-- let me give you an example of something creative that GPT-4 can already do that most people can't do. So we're still trapped in the idea of thinking that logical reasoning is the essence of intelligence when we know that being able to--

TOMASO POGGIO: --but some people.

GEOFFREY HINTON: Well--

TOMASO POGGIO: Yeah.

GEOFFREY HINTON: We know that being able to see analogies, especially remote analogies, is a very important aspect of intelligence. So I asked GPT-4, what has a compost heap got in common with an atom bomb? And GPT-4 nailed it, most people just say nothing.

DEMIS HASSABIS: What did it say?

[LAUGHTER]

GEOFFREY HINTON: It started off by saying they're very different energy scales, so on the face of it, they look to be very different. But then it got into chain reactions and how the rate at which they're generating energy increases-- their energy increases the rate at which they generate energy. So it got the idea of a chain reaction.

And the thing is, it knows about 10,000 times as much as a person, so it's going to be able to see all sorts of analogies that we can't see.

DEMIS HASSABIS: Yeah. So my feeling is on this, and starting with things like AlphaGo and obviously today's systems like Bard and GPT, they're clearly creative in some sense, like if you get them to do poetry, they're pretty amazing at poetry now. We have systems that can create great music, lots of things that we would regard as very creative. All the image stuff, text-to-image stuff.

I still think what you're asking, though, Tommy is not possible, in my opinion. I would guess. I mean, we can't categorically it because I think there's three types-- I've talked about this before probably with you, and maybe even at CBMM, about-- I think of it as three levels of creativity, and we clearly have the first two.

So first is interpolation, just averaging what you've seen to create something-- a prototypical new thing, like a new cat from all the cat images you've seen. That's the lowest level of creativity. Then there's extrapolation, which is, I think where we're at. So that's like move 37 with AlphaGo new Go strategies. New pieces of music, new pieces of poetry, and spotting analogies between things you couldn't spot as a human. And I think these systems can definitely do that.

But then there's the third level which I call like invention or out-of-the-box thinking, and that would be the equivalent of AlphaGo inventing Go. Not coming up with a good Go move, but inventing Go or inventing chess. And they can't do that. Something that's as good that we as human game aficionados would regard as classically good-- some aesthetic way as good. And they can't.

And so that's the thing that's missing. Or Picasso coming up with cubism, or a great mathematician coming up with new conjecture. But I don't believe that it's magic. I think that we will have systems that can do that, but I don't think they can do that today and there's something missing still. But I think we will be able to do that in future.

TOMASO POGGIO: So I agree with Demis. What about Ilya? Do you agree or--

ILYA SUTSKEVER: I mean, I think that the neural networks that exist today are clearly and unquestionably creative. They are not as creative as the most creative humans in history in all areas. So I think that would be a true.

TOMASO POGGIO: Yeah. I was asking about mathematics , but--

ILYA SUTSKEVER: Well, like conjecture is a bit tricky.

TOMASO POGGIO: Well--

ILYA SUTSKEVER: Like, you can already guess--

TOMASO POGGIO: --it goes exactly to what Demis said. Can you invent group theory from-- or something--

ILYA SUTSKEVER: That's different from a conjecture, I'll just say. Group theory is a little bit-- it's a pretty high bar we are talking about here.

[LAUGHTER]

Literally nothing else will remain after.

DAVID SIEGEL: I think a lot of this goes back to the benchmarking question. Even on creativity, how do you benchmark it? So I would say that one of the things that I struggle with is without benchmark-- computers forever have been able to outperform humans from almost the moment they were invented in certain tasks.

And so to really understand what's going on here-- and I'm not disagreeing with anything being said, but we really should focus on the benchmarking problem as Demis and others have pointed out.

GEOFFREY HINTON: Just a historical comment. I've lived through a long period of time when I've seen people say, neural nets will never be able to do X. The collected works of Gary Marcus is a nice history of that.

[LAUGHTER]

So I don't believe those statements anymore because for almost all the things people have said, they can now do them. And people-- proving a mathematical theorem used to be something-- well, neural networks will never do that. And people just keep moving the task, making it harder and harder.

And I completely agree with Demis. There's no reason to believe there's anything that people can do that they can't do. We may not be able to come up with-- they may not be able to come up with profound new mathematical conjectures yet, but that doesn't mean they won't be able to in 20 years' time.

ILYA SUTSKEVER: But we agree about that. We have a proof of existence. The brain is a neural network.

DEMIS HASSABIS: Yeah. I mean, unless there's something non-computable going on in the brain.

TOMASO POGGIO: Exactly. Yeah.

DEMIS HASSABIS: So, I mean, very clever ones, maybe, or very sophisticated or very evolved ones.

TOMASO POGGIO: No, my question was about the existing paradigms and transformer and large language models. And, yeah, even starting from this, not sure whether I'm invading some proprietary area here, but do you have any idea what the next architecture after transformers could be? Geoff. Geoff, you must have some idea.

GEOFFREY HINTON: If I did have an idea, I wouldn't say it in public until I had enough computation.

[LAUGHTER]

Well, at least until I had a grad student working on it.

TOMASO POGGIO: Pietro. Any idea?

PIETRO PERONA: No.

[LAUGHTER]

TOMASO POGGIO: Ilya? Yes, but you cannot speak about it. I understand.

[LAUGHTER]

All right. So--

[LAUGHTER]

AUDIENCE: So much for open science.

TOMASO POGGIO: So let me move to neuroscience. I think the question is, what would be a breakthrough in neuroscience that would be a big impact on machine learning? And I think, if we could know more about how learning is done in the brain, whether it's done by backpropagation or something else or something else is, that would be great.

I think one of the most dramatic agent in the explosive progress of neural networks has been backpropagation and gradient descent. So the question is, if that, like many people think, it's unlikely to be biologically plausible, I think could be quite interesting to know how the brain does it, and that may have the potential to have an impact on AI as well.

But you may think that there are other potential breakthrough in neuroscience that may have an impact in machine learning. Do you have any idea?

GEOFFREY HINTON: Well, I think it's fairly clear that the brain isn't going to be doing backpropagation through time. That seems very unlikely. All the theories of how the brain does backpropagation that I know of are for doing backpropagation through multiple cortical areas.

It's also fairly clear that these big language models, and the multimodal ones, too, are storing far more information per connection than the brain. Now it may just be because they've got far more experience and we will be able to get much more if we have much more experience.

But I suspect now-- I've always thought the brain must be doing some form of backpropagation, but I now suspect it may be doing something dumber. But if I could get one question answered by GPT-20, it would be, does the brain implement some form of backpropagation?

TOMASO POGGIO: This was great. Thank you all.

[APPLAUSE]

DEMIS HASSABIS: Thanks. Thanks, everyone.