Dwarkesh Podcast #61 - Shane Legg (DeepMind Founder) - 2028 AGI, Superhuman Alignment, New Architectures

Published - October 26, 2023

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
Today I have the pleasure of interviewing Shane Legg, who is the founder and the Chief AGI scientist of Google DeepMind. Shane, welcome to the podcast.

Shane Legg
Thank you. It's a pleasure being here.

Dwarkesh Patel
First question. How do we measure progress towards AGI concretely? We have these loss numbers and we can see how the loss improves from one model to another, but it's just a number. How do we interpret this? How do we see how much progress we're actually making?

Shane Legg
That’s a hard question. AGI by its definition is about generality. It's not about doing a specific thing. It's much easier to measure performance when you have a very specific thing in mind because you can construct a test around that.

Maybe I should first explain what I mean by AGI because there are a few different notions around it. When I say AGI, I mean a machine that can do the sorts of cognitive things that people can typically do, possibly more. To be an AGI that's the bar you need to meet.

So if we want to test whether we're meeting the threshold or we're getting close to the threshold, what we actually need is a lot of different kinds of measurements and tests that span the breadth of all the sorts of cognitive tasks that people can do and then to have a sense of what human performance is on these sorts of tasks. That then allows us to judge whether or not we're there.

It's difficult because you'll never have a complete set of everything that people can do because it's such a large set. But I think that if you ever get to the point where you have a pretty good range of tests of all sorts of cognitive things that we can do, and you have an AI system which can meet human performance and all those things and then even with effort, you can't actually come up with new examples of cognitive tasks where the machine is below human performance then at that point, you have an AGI.

It may be conceptually possible that there is something that the machine can't do that people can do but if you can't find it with some effort, then for all practical purposes, you have an AGI.

Dwarkesh Patel
Let's get more concrete. We measure the performance of these large language models on MMLU and other benchmarks. What is missing from the benchmarks we use currently? What aspect of human cognition do they not measure adequately?

Shane Legg
Another hard question. These are quite big areas. They don't measure things like understanding streaming video, for example, because these are language models and people can do things like understanding streaming video.

They don't do things like episodic memory. Humans have what we call episodic memory. We have a working memory, which are things that have happened quite recently, and then we have a cortical memory, things that are sort of being in our cortex, but there's also a system in between, which is episodic memory, which is the hippocampus. It is about learning specific things very, very rapidly. So if you remember some of the things I say to you tomorrow, that'll be your episodic memory hippocampus.

Our models don't really have that kind of thing and we don't really test for that kind of thing. We just sort of try to make the context windows, which is more like working memory, longer and longer to sort of compensate for this.

But it is a difficult question because the generality of human intelligence is very, very broad. So you really have to start going into the weeds of trying to find if there's specific types of things that are missing from existing benchmarks or different categories of benchmarks that don't currently exist or something.

Dwarkesh Patel
The thing you're referring to with episodic memory, would it be fair to call that sample efficiency or is that a different thing?

Shane Legg
It's very much related to sample efficiency. It's one of the things that enables humans to be very sample efficient. Large language models have a certain kind of sample efficiency because when something's in their context window, that biases the distribution to behave in a different way and so that's a very rapid kind of learning. There are multiple kinds of learning and the existing systems have some of them, but not others. It's a little bit complicated.

Dwarkesh Patel
Is this kind of memory, what we call sample efficiency, a fatal flaw of these deep learning models that it just takes trillions of tokens, a magnitude more than what any human will see throughout their lifetime or is this something that will be solved over time?

Shane Legg
The models can learn things immediately when it's in the context window and then they have this longer process when you actually train the base model and that's when they're learning over trillions of tokens. But they miss something in the middle. That's sort of what I'm getting at here.

I don't think it's a fundamental limitation. I think what's happened with large language models is something fundamental has changed. We know how to build models now that have some degree of understanding of what's going on. And that did not exist in the past. And because we've got a scalable way to do this now, that unlocks lots and lots of lots of new things.

Now we can look at things which are missing, such as this sort of episodic memory type thing, and we can then start to imagine ways to address that. My feeling is that there are relatively clear paths forward now to address most of the shortcomings we see in the existing models, whether it's about delusions, factuality, the type of memory and learning that they have, or understanding video, or all sorts of things like that. I don't see any big blockers. I don't see big walls in front of us. I just see that there's more research and work and all these things will improve and probably be adequately solved.

Dwarkesh Patel
Going back to the original question of how do you measure when human level AI has arrived or has gone beyond it. As you mentioned, there's these other sorts of benchmarks you can use and other sorts of traits, but concretely, what would it have to do for you to be like, “Okay, we've reached human level.”

Would it have to beat Minecraft from start to finish? Would it have to get 100% on MMLU? What would it have to do?

Shane Legg
There is no one thing that would do it, because I think that's the nature of it. It's about general intelligence. So I'd have to make sure it could do lots and lots of different things and it didn't have a gap.

We already have systems that can do very impressive categories of things to human level or even beyond. I would want a whole suite of tests that I felt was very comprehensive and then furthermore, when people come in and say, “Okay, so it's passing a big suite of tests, let's try to find examples. Let's take an adversarial approach to this. Let's deliberately try to find examples where people can clearly, typically do this, but the machine fails.” And when those people cannot succeed, I'll go, “Okay, we're probably there.”

Dwarkesh Patel
A lot of your earlier research, at least the ones I could find, emphasized that AI should be able to manipulate and succeed in a variety of open-ended environments. It almost sounds like a video game. Is that where your head is still at now, or do you think about it differently?

Shane Legg
It’s evolved a bit. When I did my thesis work around universal intelligence, I was trying to come up with an extremely universal, general, mathematically clean framework for defining and measuring intelligence. I think there were aspects of that that were successful. I think in my own mind, it clarified the nature of intelligence as being able to perform well in lots of different domains and different tasks and so on. It's about that sort of capability of performance and the breadth of performance. I found that was quite helpful and enlightening.

There was always the issue of the reference machine. In the framework, you have a weighting of things according to the complexity. It's like an Occam's razor type of thing, where you weight tasks and environments which are simpler, more highly. You’ve got a countable space of semi-computable environments. And that Kolmogorov complexity measure has something built into it, which is called a reference machine. And that's a free parameter. So that means that the intelligence measure has a free parameter in it and as you change that free parameter, it changes the weighting and the distribution over the space of all the different tasks and environments. This is sort of an unresolved part of the whole problem. So what reference machine should we ideally use? There's no universal reference machine. People will usually put a Turing machine in there, but there are many kinds of different machines.

Given that it's a free parameter, I think the most natural thing to do is to think about what's meaningful to us in terms of intelligence. I think human intelligence is meaningful to us in the environment that we live in. We know what human intelligence is. We are human too. We interact with other people who have human intelligence. We know that human intelligence is possible, obviously, because it exists in the world. We know that human intelligence is very, very powerful because it's affected the world profoundly in countless ways. And we know if human level intelligence was achieved, that would be economically transformative because the types of cognitive tasks people do in the economy could be done by machines then. And it would be philosophically important because this is sort of how we often think about intelligence. Historically it would be a key point.

So I think that human intelligence in a human-like environment is quite a natural sort of reference point. You could imagine setting your reference machine to be such that it emphasizes the kinds of environments that we live in as opposed to some abstract mathematical environment. And so that's how I've kind of gone on this journey of — “Let's try to define a completely universal, clean, mathematical notion of intelligence" to "Well, it's got a free parameter. "

One way of thinking about it is to think more concretely about human intelligence and build machines that can match human intelligence. Because we understand what that is and we know that that is a very powerful thing. It has economic, philosophical and historical importance.

The other aspect of course is that, in this pure formulation of Kolmogorov complexity, it's actually not computable. I also knew that there was a limitation at the time but it was an effort to just theoretically come up with a clean definition. I think we can sort of get there, but we have this issue of a reference machine, which is unspecified.

Dwarkesh Patel
Before we move on, I do want to ask a question on the original point you made on LLMs needing episodic memory. You said that these are problems that we can solve and these are not fundamental impediments.

But when you say that, do you think they will just be solved by scale or do each of these need a fine-grained specific solution that is architectural in nature?

Shane Legg
I think it'll be architectural in nature because the current architectures don't really have what you need to do this. They basically have a context window, which is very, very fluid, of course, and they have the weights, which things get baked into very slowly. So to my mind, that feels like working memory, which is like the activations in your brain, and then the weights are like the synapses in your cortex.

Now, the brain separates these things out. It has a separate mechanism for rapidly learning specific information because that's a different type of optimization problem compared to slowly learning deep generalities. There's a tension between the two but you want to be able to do both. You want to be able to hear someone's name and remember it the next day. And you also want to be able to integrate information over a lifetime so you start to see deeper patterns in the world.

These are quite different optimization targets, different processes, but a comprehensive system should be able to do both. And so I think it's conceivable you could build one system that does both, but you can also see that because they're quite different things, it makes sense for them to be done differently. I think that's why the brain does it separately.

Dwarkesh Patel
I'm curious about how concretely you think that would be achieved. DeepMind has been working on these domain specific reinforcement learning type setups: AlphaFold, AlphaCode and so on. How does that fit into what you see as a path to AGI? Have these just been orthogonal domain specific models or do they feed into the eventual AGI?

Shane Legg
Things like AlphaFold are not really feeding into AGI. We may learn things in the process that may end up being relevant, but I don't see them as likely being on the path to AGI. But we're a big group. We've got hundreds and hundreds and hundreds of PhDs working on lots of different projects. When we find what we see as opportunities to do something significant like AlphaFold, we'll go and do it. It's not like we only do AGI type work. We work on fusion reactors and various things in sustainability, energy. We've got people looking at satellite images of deforestation. We have people looking at weather forecasting. We've got tons of people working on lots of things.

Dwarkesh Patel
On the point you made earlier about the reference machine as human intelligence. It's interesting because one of the things you mentioned in your 2008 thesis about how you would measure intelligence was — You said you could do a compression test and you could see if it fills in words and a sample of text and that could measure intelligence. And funnily enough, that's basically how the LLMs are trained.

At the time, did it stick out to you as an especially fruitful thing to train for?

Shane Legg
Well, yeah. In the sense what's happened is actually very aligned with what I wrote about in my thesis. The ideas from Marcus Hutter with AIXI, where you take Solomonoff induction, which is this incomputable but theoretically very elegant and extremely sample efficient prediction system, and then once you have that, you can build a general agent on top of it by basically adding search and reinforcement signal. That's what you do with AIXI.

But what that sort of tells you is that if you have a fantastically good sequence predictor, some approximation of Solomonoff induction, then going from that to a very powerful, very general AGI system is just sort of another step. You've actually solved a lot of the problem already.

And I think that's what we're seeing today actually, that these incredibly powerful foundation models are incredibly good sequence predictors that are compressing the world based on all this data. And then you will be able to extend these in different ways and build very, very powerful agents.

Dwarkesh Patel
Let me ask you more about that.

Richard Sutton's Bitter Lesson essay says that there's two things you can scale, search and learning. I guess you could say that LLMs are about the learning aspect. The search stuff, which you worked on throughout your career, where you have an agent that is interacting with this environment, is that the direction that needs to be explored again? Or is that something that needs to be added to LLMs where they can actually interact with their data or the world or in some way?

Shane Legg
Yeah, I think that's on the right track. These foundation models are world models of a kind and to do really creative problem solving, you need to start searching. If I think about something like AlphaGo and the famous Move 37, where did that come from? Did that come from all its data that it's seen of human games or something like that? No, it didn't. It came from it identifying a move as being quite unlikely, but plausible. And then via a process of search, coming to understand that it was actually a very, very good move.

So to get real creativity, you need to search through spaces of possibilities and find these hidden gems. That's what creativity is. Current language models don't really do that. They really are mimicking the data. They are mimicking all the human ingenuity and everything, which they have seen from all this data that's coming from the internet that's originally derived from humans.

These models can blend things. They can do Harry Potter in the style of a Kanye West rap or something, even though it's never happened, they can blend things together. But if you want a system that can go truly beyond that and not just generalize in novel ways and do something that's truly creative, that is not just a blending of existing things, that requires searching through a space of possibilities and finding these hidden gems that are hidden away in there somewhere. And that requires search. So I don't think we'll see systems that truly step beyond their training data until we have powerful search in the process.

Dwarkesh Patel
There are rumors that Google DeepMind is training newer models, and you don't have to comment on those specifically, but when you do that, if it's the case that something like search is required to go to the next level, are you training in a completely different way than how GPT-4 or other transformers are trained?

Shane Legg
And I can't say much about how we're training. I think it's fair to say we're roughly doing the sorts of scaling and training that you see many people in the field doing but we have our own take on it and our own different tricks and techniques.

Dwarkesh Patel
Okay, maybe we'll come back to it and get another answer on that.

Let's talk about alignment briefly. What will it take to align human level and superhuman AIs?

It's interesting because the sorts of reinforcement learning and self-play kinds of setups that are popular now, like Constitution AI or RLHF, DeepMind obviously has expertise in it for decades longer. I'm curious what you think of the current landscape and how DeepMind pursues that problem of safety towards human level models.

Shane Legg
Do you want to know about what we're currently doing or do you want me to have a stab at what I think needs to be done?

Dwarkesh Patel
Needs to be done.

Shane Legg
Currently we're doing lots of things. We're doing interpretability. We're doing our process supervision. We're doing red teaming. We're doing evaluation for dangerous capabilities. We're doing work on institutions and governance and tons of stuff, right?

Anyway, what do I think needs to be done? I think that powerful machine learning, powerful AGI, is coming in some time and if the system is really capable, really intelligent, really powerful, trying to somehow contain it or limit it is probably not a winning strategy because these systems ultimately will be very, very capable. So what you have to do is you have to align it. You have to get it such that it's fundamentally a highly ethical value aligned system from the get go. How do you do that?

Maybe this is slightly naive, but this is my take on it — How do people do it? If you have a really difficult ethical decision in front of you, what do you do? You don't just do the first thing that comes to mind, because there could be a lot of emotions involved in other things. It's a difficult problem.

What you have to do is to calm yourself down. You've got to sit down and you've got to think about it. You've got to think, “Well, okay, what could I do?” I could do this. I could do this. I could do this. If I do each of these things, what will happen? So that requires a model of the world. And then you have to think about ethically, how do I view each of these different actions and the possibilities and what might happen from it? What is the right thing to do? And as you think about all the different possibilities and your actions and what can follow from them and how it aligns with your values and your ethics, you can then come to some conclusion of what is really the best choice that you should be making if you want to be really ethical about this.

I think AI systems need to essentially do the same thing. When you sample from a foundation model at the moment, it's blurting out the first thing. It's like System 1, if you like, from psychology, from Kahneman et al. That's not good enough.

And if we do RLHF without human feedback (RLAIF), Constitutional AI tries to do that sort of thing, you're trying to fix the underlying System 1 in a sense. That can shift the distribution and that can be very helpful but it's a very high dimensional distribution and you're sort of poking it in a whole lot of points. So it's not likely to be a very robust solution. It's like trying to train yourself out of a bad habit. You can sort of do it eventually. But what you need to do is you need to have a System 2. You need the system to not just sample from the model. You need the system to go, “Okay, I'm going to reason this through. I'm going to do step by step reasoning. What are the options in front of me? I'm going to use my world model now and I'm going to use a good world model to understand what's likely to happen from each of these options.” And then reason about each of these from an ethical perspective.

So you need a system which has a deep understanding of the world, a good world model, and has a good understanding of people, and has a good understanding of ethics, and it has robust and very reliable reasoning. And then you set it up in such a way that it applies this reasoning and this understanding of ethics to analyze the different options which are in front of it and then execute on which is the most ethical way forward.

Dwarkesh Patel
But when a lot of people think about the fundamental alignment problem, the worry is not that it's not going to have a world model to understand the effects of its actions, the worry is that the effects it cares about are not the ones we will care about. So even if you improve its system-2 thinking to do better planning, the fundamental problem is — We have these really nuanced values about what we want. How do we communicate those values and make sure they're reinforced in the AI?

Shane Legg
It needs not just a good model of the world, but it needs a really good understanding of ethics. And we need to communicate to the system what ethics and values it should be following.

Dwarkesh Patel
And how do we do that in a way that we can be confident that a super human level model will preserve those values or have learned them in the first place?

Shane Legg
It should preserve them because if it's making all its decisions based on a good understanding of ethics and values, and it's consistent in doing this, it shouldn't take actions which undermine that. That would be inconsistent.

Dwarkesh Patel
Right, so then how do we get to the point where it has learned them in the first place?

Shane Legg
Yeah, that's the challenge. We need to have systems. The way I think about it is this: to have a profoundly ethical AI system, it also has to be very, very capable. It needs a really good world model, a really good understanding of ethics, and it needs really good reasoning. Because if you don't have any of those things, how can you possibly be consistently profoundly ethical? You can't. So we actually need better reasoning, better understanding of the world, and better understanding of ethics in our systems.

Dwarkesh Patel
It seems to me that the former two would just come along for the ride as these models get more powerful.

Shane Legg
Yeah. That's a nice property because it's actually a capabilities thing to some extent.

Dwarkesh Patel
But if the third one, the ethical model, is a bottleneck, or if it’s a thing that doesn't come along with the AI itself, what is the actual technique to make sure that that happens?

Shane Legg
First of all, we should train the system on ethics generally so that it understands human ethics well. There's a lot of lectures and papers and books and all sorts of things. We need to make sure it understands humans ethics well, at least as well as a very good ethicist because that's important.

And we then need to decide, of this general understanding of ethics, what do we want the system to actually value and what sort of ethics do we want it to apply? Now, that's not a technical problem. That's a problem for society and ethicists and so on to come up with.

I'm not sure there's such a thing as optimal ethics but I'm pretty sure that it's possible to come up with a set of ethics, which is much better than what the so-called doomers are worried about in terms of the behavior of these AGI systems. And then what you do is you engineer the system to actually follow these things so that every time it makes a decision, it does an analysis using a deep understanding of the world and of ethics and very robust and precise reasoning to do an ethical analysis of what it's doing.

And of course, we would want lots of other things. We would want people checking these processes of reasoning. We'd want people verifying that it's behaving itself in terms of how it reaches these conclusions.

Dwarkesh Patel
But I still feel like I don't understand how that fundamental problem of making sure it follows that ethic works. Because presumably, it has read Mao's books so it understands Maoist ethics and understands all these other ethics. How do we make sure the ethic that ethicists say is the one is what it ends up following and not the other ones it understands?

Shane Legg
Right. So you have to specify to the system, these are ethical principles that you should follow.

Dwarkesh Patel
And how do we make sure it does that?

Shane Legg
We have to check it as it's doing it. We have to assure ourselves that it is consistently following these ethical principles at least as well as a group of human experts.

Dwarkesh Patel
Are you worried that if you do it the default way, which is just reinforcing it whenever it seems to be following them, you could be training deception as well?

Shane Legg
Reinforcement does have some dangerous aspects to it. I think it's actually more robust to check the process of reasoning and check its understanding of ethics. To reassure ourselves that the system has a really good understanding of ethics, it should be grilled for some time to try to really pull apart its understanding and make sure it is very robust.

And also, if it's deployed, we should have people constantly looking at the decisions it's making and the reasoning process that goes into those decisions to try to make sure that it is correctly reasoning about these types of things.

Dwarkesh Patel
Do you have some sort of framework for that at Google DeepMind?

Shane Legg
This is not so much a Google DeepMind perspective on this. This is my take on how I think we need to do this kind of thing. There are many different views within and there are different variants on these sorts of ideas as well.

Dwarkesh Patel
So then do you personally think there needs to be some sort of framework for as you arrive at certain capabilities, these are the concrete safety benchmarks that you must have instated at this point, or you should pause or slow down?

Shane Legg
I think that's a sensible thing to do but it's actually quite hard to do. There are some people thinking about that. I know Anthropic has put out some things like that. We were thinking about similar things but actually putting concrete things down is quite a hard thing to do. I think it's an important problem and I certainly encourage people to work on it.

Dwarkesh Patel
It's interesting because you have these blog posts that you wrote when you started DeepMind, back in 2008, where the motivation was to accelerate safety.

On net, what do you think the impact of DeepMind has been on safety versus capabilities?

Shane Legg
Ooh, interesting. I don't know. It's hard to judge, actually.

I've been worried about AGI safety for a long time, well before DeepMind. But it was always really hard to hire people to work on AGI safety, particularly in the early days. Back in 2013 or so, we had our first hire and he only agreed to do it part-time because he didn't want to drop all the capabilities work because of the impact it could have on his career. And this was someone who had already previously been publishing in AGI safety.

I don't know. It's hard to know what is the counterfactual if we weren't there doing it. We have been a group that has talked about this openly. I've talked about the importance of it on many occasions. We've been hiring people to work on these topics. I know a lot of other people in the area and I've talked to them over many, many years. I've known Dario since 2005 or something and we've talked on and off about AGI safety and so on.

The impact that DeepMind has had: I guess we were the first AGI company and as the first AGI company, we always had an AGI safety group. We've been publishing papers on this for many years. I think that's lent some credibility to the area of AGI safety. AGI was a fringe term not that long ago. I hope that creates some space for people.

Dwarkesh Patel
Where do you think AI progress itself would have been without DeepMind?

This is not just a point that people make about DeepMind. I think this is a general point people make about OpenAI and Anthropic as well, that these people went into the business to accelerate safety and the net effect might have been to accelerate capabilities far more.

Shane Legg
Right, right. I think we have accelerated capabilities, but again, the counterfactuals are quite difficult. We didn't do ImageNet, for example, and ImageNet was very influential in attracting investment to the field. We did do AlphaGo, and that changed some people's minds. But, the community is a lot bigger than just DeepMind.

If you went back more than five years in the future, we were able to do bigger projects with bigger teams and take on more ambitious things than a lot of the smaller academic groups, right? And so the sort of nature of the type of work we could do was a bit different. And that affected the dynamics in some ways.

But, the community is much, much bigger than DeepMind. There are a number of other players with significant resources. Maybe we've sped things up a bit, but I think a lot of these things would have happened before too long anyway. Often good ideas are in the air, and as a researcher, when you're about to publish something, you see somebody else has got a very similar idea coming out with some good results. Often it's kind of like the time is right for things. So I find it very hard to reason about the counterfactuals there.

Dwarkesh Patel
Speaking of the early years, it's really interesting that in 2011, you had a blog post where you said — "I've decided to once again leave my prediction for when human level AGI will arrive unchanged. That is, I give it a log-normal distribution with a mean of 2028 and a mode of 2025, under the assumption that nothing crazy happens like a nuclear war."

This is before deep learning, this is when nobody's talking about AI, and it turns out that if the trends continue, this is not an unreasonable prediction.

How did you have that accurate of an estimate before all these trends came into effect?

Shane Legg
First I'd say it's not before deep learning. Deep learning was getting started around 2008.

Dwarkesh Patel
Oh, sorry. I meant to say before ImageNet.

Shane Legg
Before ImageNet? Yeah, that was 2012.

I first formed those beliefs around 2001 after reading Ray Kurzweil's The Age of Spiritual Machines. There were two really important points in his book that I came to believe as true. One is that computational power would grow exponentially for at least a few decades. And that the quantity of data in the world would grow exponentially for a few decades. And when you have exponentially increasing quantities of computation and data, then the value of highly scalable algorithms gets higher and higher. There's a lot of incentive to make a more scalable algorithm to harness all this computing data. So I thought it would be very likely that we'll start to discover scalable algorithms to do this. And then there's a positive feedback between all these things, because if your algorithm gets better at harnessing computing data, then the value of the data and the compute goes up because it can be more effectively used. And that drives more investment in these areas. If your compute performance goes up, then the value of the data goes up because you can utilize more data. So there are positive feedback loops between all these things. That was the first thing.

And then the second thing was just looking at the trends. If the scalable algorithms were to be discovered, then during the 2020s, it should be possible to start training models on significantly more data than a human would experience in a lifetime. And I figured that that would be a time where big things would start to happen that would eventually unlock AGI. So that was my reasoning process. And I think we're now at that first part. I think we can start training models now with the scale of the data that is beyond what a human can experience in a lifetime. So I think this is the first unlocking step.

And so, yeah, I think there's a 50% chance that we have AGI by 2028. Now, it's just a 50% chance. I'm sure what's going to happen is we're going to get to 2029 and someone's going to say, “Shane, you were wrong.” Come on, I said 50% chance.

I think it's entirely plausible but I'm not going to be surprised if it doesn't happen by then. You often hit unexpected problems in research and science and sometimes things take longer than you expect.

Dwarkesh Patel
If we're in 2029 and it hasn't happened yet, if there was a problem that caused it, what would be the most likely reason for that?

Shane Legg
I don't know. At the moment, it looks to me like all the problems are likely solvable with a number of years of research. That's my current sense.

Dwarkesh Patel
And what does the time from here to 2028 look like if 2028 ends up being the year?

Is it just that we have trillions of dollars of economic impact in the meantime and the world gets crazy or what happens?

Shane Legg
I think you'll see the existing models maturing. They'll be less delusional, much more factual. They'll be more up to date on what's currently going on when they answer questions. They'll become multimodal, much more than they currently are. And this will just make them much more useful.

So I think probably what we'll see more than anything is just loads of great applications for the coming years. There can be some misuse cases as well. I'm sure somebody will come up with something to do with these models that is quite unhelpful. But my expectation for the coming years is mostly a positive one. We'll see all kinds of really impressive, really amazing applications for the coming years.

Dwarkesh Patel
And on the safety point, you mentioned these different research directions that are out there and that you are doing internally in DeepMind as well. Interpretability, RLAIF and so on. Which are you most optimistic about?

Shane Legg
Oooh. I don't know. I don't want to pick favorites. It's hard picking favorites. I know the people working on all these areas. I think things of the sort of system 2 flavor. There's work we have going on that Geoffrey Irving leads called Deliberative Dialogue, which has the System 2 flavor where a sort of debate takes place about the actions that an agent could take or what's the correct answer to something like this. And people then can sort of review these debates and so on. And they use these AI algorithms to help them judge the correct outcomes and so on. And so this is sort of meant to be a way in which to try to scale the alignment to increasingly powerful systems. I think things of that kind of flavor have quite a lot of promise in my opinion, but that's kind of quite a broad category. There are many different topics within that.

Dwarkesh Patel
That's interesting. So you mentioned two areas in which LLMs needs to improve. One is the episodic memory and the other is the System 2 thinking. Are those two related or are they two separate drawbacks?

Shane Legg
I think they're fairly separate, but they can be somewhat related. You can learn different ways of thinking through problems and actually learn about this rapidly using your episodic memory. All these different systems and subsystems interact so they're never completely separate. But I think conceptually you can probably think of them as quite separate things.

I think delusions and factuality is another area that's going to be quite important and particularly important in lots of applications. If you want a model that writes creative poetry, then that's fine because you want to be able to be very free to suggest all kinds of possibilities and so on. You're not really constrained by a specific reality. Whereas if you want something that's in a particular application, normally you have to be quite concrete about what's currently going on and what is true and what is not true and so on. And models are a little bit sort of freewheeling when it comes to truth and creativity at the moment. And that I think limits their applications in many ways.

Dwarkesh Patel
The final question is this. You've been in this field for over a decade, much longer than many others, and you've seen different landmarks like ImageNet and Transformers. What do you think the next landmark will look like?

Shane Legg
I think the next landmark that people will think back to and remember is going much more fully multimodal. That will open out the sort of understanding that you see in language models into a much larger space of possibilities. And when people think back, they'll think about, “Oh, those old fashioned models, they just did like chat, they just did text.” It just felt like a very narrow thing whereas now they understand when you talk to them and they understand images and pictures and video and you can show them things or things like that. And they will have much more understanding of what's going on. And it'll feel like the system's kind of opened up into the world in a much more powerful way.

Dwarkesh Patel
Do you mind if I ask a follow-up on that? ChatGPT just released their multimodal feature and you, in DeepMind, you had the Gato paper, where you have this one model where you can throw images, video games and even actions in there. So far it doesn't seem to have percolated as much as ChatGPT initially from GPT3 or something.

What explains that? Is it just that people haven't learned to use multimodality? They're not powerful enough yet?

Shane Legg
I think it's early days. I think you will see understanding images and things more and more. But I think it's early days in this transition is when you start really digesting a lot of video and

other things like that, that the systems will start having a much more grounded understanding of the world and all kinds of other aspects. And then when that works well, that will open up naturally lots and lots of new applications and all sorts of new possibilities because you're not confined to text chat anymore.

Dwarkesh Patel
New avenues of training data as well, right?

Shane Legg
Yeah, new training data and all kinds of different applications that aren't just purely textual anymore. And what are those applications? Well, probably a lot of them we can't even imagine at the moment because there are just so many possibilities once you can start dealing with all sorts of different modalities in a consistent way.

Dwarkesh Patel
Awesome. I think that's an actionable place to leave it off. Thank you so much for coming on the podcast Shane.

Shane Legg
Thank you.