Dwarkesh Podcast #70 - Demis Hassabis - Scaling, Superhuman AIs, AlphaZero atop LLMs, Rogue Nations Threat

Published - February 28, 2024

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
Today it is a true honor to speak with Demis Hassabis, who is the CEO of DeepMind. Demis, welcome to the podcast.

Demis Hassabis
Thanks for having me.

Dwarkesh Patel
First question, given your neuroscience background, how do you think about intelligence? Specifically, do you think it's one higher-level general reasoning circuit, or do you think it's thousands of independent subskills and heuristics?

Demis Hassabis
It's interesting because intelligence is so broad and what we use it for is so generally applicable. I think that suggests there must be high-level common algorithmic themes around how the brain processes the world around us. Of course, there are specialized parts of the brain that do specific things, but I think there are probably some underlying principles that underpin all of that.

Dwarkesh Patel
How do you make sense of the fact that in these LLMs, when you give them a lot of data in any specific domain, they tend to get asymmetrically better in that domain. Wouldn't we expect a general improvement across all the different areas?

Demis Hassabis
First of all, I think you do sometimes get surprising improvement in other domains when you improve in a specific domain. For example, when these large models improve at coding, that can actually improve their general reasoning. So there is evidence of some transfer although we would like a lot more evidence of that. But that's how the human brain learns too. If we experience and practice a lot of things like chess, creative writing, or whatever, we also tend to specialize and get better at that specific thing even though we're using general learning techniques and general learning systems in order to get good at that domain.

Dwarkesh Patel
What's been the most surprising example of this kind of transfer for you? Will you see language and code, or images and text?

Demis Hassabis
I'm hoping we're going to see a lot more of this kind of transfer, but I think things like getting better at coding and math, and then generally improving your reasoning. That is how it

works with us as human learners. But I think it's interesting seeing that in these artificial systems.

Dwarkesh Patel
And can you see the sort of mechanistic way, in the language and code example, in which you've found the place in a neural network that's getting better with both the language and the code? Or is that too far down the weeds?

Demis Hassabis
I don't think our analysis techniques are quite sophisticated enough to be able to hone in on that. I think that's actually one of the areas where a lot more research needs to be done, the kind of mechanistic analysis of the representations that these systems build up. I sometimes like to call it virtual brain analytics. In a way, it's a bit like doing fMRI, or single-cell recording from a real brain. What are the analogous analysis techniques for these artificial minds? There's a lot of great work going on in this sort of stuff. People like Chris Olah, I really like his work. I think a lot of computational neuroscience techniques can be brought to bear on analyzing the current systems we're building. In fact, I try to encourage a lot of my computational neuroscience friends to start thinking in that direction and applying their know-how to the large models.

Dwarkesh Patel
What do other AI researchers not understand about human intelligence that you have some sort of insight on, given your neuroscience background?

Demis Hassabis
I think neuroscience has added a lot, if you look at the last 10-20 years that we've been at it. I've been thinking about this for 30+ years. In the earlier days of the new wave of AI, neuroscience was providing a lot of interesting directional clues, things like reinforcement learning and combining that with deep learning. Some of our pioneering work we did there were things like experience replay and even the notion of attention, which has become super important. A lot of those original inspirations came from some understanding about how the brain works, although not the exact specifics of course. One is an engineered system and the other one's a natural system. It's not so much about a one-to-one mapping of a specific algorithm, but more so inspirational direction. Maybe it's some ideas for architecture, or algorithmic ideas, or representational ideas. The brain is an existence proof that general intelligence is possible at all. I think the history of human endeavors has been such that once you know something's possible it's easier to push hard in that direction, because you know it's a question of effort, a question of when and not if. That allows you to make progress a lot more quickly. So I think neuroscience has inspired a lot of the thinking, at least in a soft way, behind where we are today. As for going forward, I think there's still a lot of interesting things to be resolved around planning. How does the brain construct the right world models? I studied how the brain does imagination, or you can think of it as

mental simulation. How do we create very rich visual spatial simulations of the world in order for us to plan better?

Dwarkesh Patel
Actually, I'm curious how you think that will interface with LLMs. Obviously, DeepMind is at the frontier and has been for many years with systems like AlphaZero and so forth, having these agents which can think through different steps to get to an end outcome. Is there a path for LLMs to have this tree search kind of thing on top of them? How do you think about this?

Demis Hassabis
I think that's a super promising direction. We've got to carry on improving the large models. We've got to carry on making them more and more accurate predictors of the world, making them more and more reliable world models. That's clearly a necessary, but probably insufficient component of an AGI system. On top of that, we're working on things like AlphaZero-like planning mechanisms on top that make use of that model in order to make concrete plans to achieve certain goals in the world. Perhaps chaining thought, lines of reasoning, together and using search to explore massive spaces of possibility. I think that's kind of missing from our current large models.

Dwarkesh Patel
How do you get past the immense amount of compute that these approaches tend to require? Even the AlphaGo system was a pretty expensive system because you sort of had to run an LLM on each node of the tree. How do you anticipate that'll get made more efficient?

Demis Hassabis
One thing is Moore's law tends to help. Over every year more computation comes in. But we focus a lot on sample-efficient methods and reusing existing data, things like experience replay and also just looking at more efficient ways. The better your world model is, the more efficient your search can be. One example I always give is AlphaZero, our system to play Go and chess and any game. It's stronger than human world champion level in all these games and it uses a lot less search than a brute force method like Deep Blue to play chess. One of these traditional Stockfish or Deep Blue systems would maybe look at millions of possible moves for every decision it's going to make. AlphaZero and AlphaGo may look at around tens of thousands of possible positions in order to make a decision about what to move next. A human grandmaster or world champion probably only looks at a few hundred moves, even the top ones, in order to make their very good decision about what to play next. So that suggests that the brute force systems don't have any real model other than the heuristics about the game. AlphaGo has quite a decent model but the top human players have a much richer, much more accurate model of Go or chess. That allows them to make world-class decisions on a very small amount of search. So I think there's a sort of trade-off there. If you

improve the models, then I think your search can be more efficient and therefore you can get further with your search.

Dwarkesh Patel
I have two questions based on that. With AlphaGo, you had a very concrete win condition: at the end of the day, do I win this game of Go or not? You can reinforce on that. When you're thinking of an LLM putting out thought, do you think there will be this ability to discriminate in the end, whether that was a good thing to reward or not?

Demis Hassabis
Of course that's why we pioneered, and what DeepMind is sort of famous for, using games as a proving ground. That's partly because it's efficient to research in that domain. The other reason is, obviously, it's extremely easy to specify a reward function. Winning the game or improving the score, something like that is built into most games. So that is one of the challenges of real-world systems. How does one define the right objective function, the right reward function, and the right goals? How does one specify them in a general way, but specific enough that one actually points the system in the right direction? For real-world problems, that can be a lot harder. But actually, if you think about it in even scientific problems, there are usually ways that you can specify the goal that you're after.

Dwarkesh Patel
When you think about human intelligence, you were just saying that humans thinking about these thoughts are just super sample-efficient. Einstein coming up with relativity, right? There's thousands of possible permutations of the equations. Do you think it's also this sense of different heuristics like, "I'm going to try out this approach instead of this"? Or is it a totally different way of approaching and coming up with that solution than what AlphaGo does to plan the next move?

Demis Hassabis
I think it's different because our brains are not built for doing Monte Carlo tree search. It's just not the way our organic brains work. I think that people like Einstein, in order to compensate for that, have used their intuition—and maybe we can come to what intuition is—and their knowledge and their experience to build in Einstein's case, extremely accurate models of physics that include mental simulations. If you read about Einstein and how he came up with things, he used to visualize and really feel what these physical systems should be like, not just the mathematics of it. He had a really intuitive feel for what they would be like in reality. That allowed him to think these thoughts that were very outlandish at the time. So I think that that gets to the sophistication of the world models that we're building. Imagine your world model can get you to a certain node in a tree that you're searching, and then you just do a little bit of search around that leaf node and that gets you to these original places. Obviously, if your model and your judgment on that model is very, very good, then you can pick which leaf nodes you should expand with search much more accurately.

So overall, you therefore do a lot less search. I mean, there's no way that any human could do a kind of brute force search over any kind of significant space.

Dwarkesh Patel
A big open question right now is whether RL will allow these models to use the self-play synthetic data to get over data bottlenecks. It sounds like you're optimistic about this?

Demis Hassabis
I'm very optimistic about that. First of all, there's still a lot more data that can be used, especially if one views multimodal and video and these kinds of things. Obviously, society is adding more data all the time to the Internet and things like that. I think that there's a lot of scope for creating synthetic data. We're looking at that in different ways, partly through simulation, using very realistic game environments, for example, to generate realistic data, but also self-play. That's where systems interact with each other or converse with each other. It worked very well for us with AlphaGo and AlphaZero where we got the systems to play against each other and actually learn from each other's mistakes and build up a knowledge base that way. I think there are some good analogies for that. It's a little bit more complicated to build a general kind of world data.

Dwarkesh Patel
How do you get to the point with these models where the synthetic data they're outputting on the self-play they're doing is not just more of what's already in their data set, but something they haven't seen before? To actually improve the abilities.

Demis Hassabis
I think there's a whole science needed there. I think we're still in the nascent stage of this, of data curation and data analysis and actually analyzing the holes that you have in your data distribution. This is important for things like fairness and bias and other stuff. To remove that from the system is to really make sure that your data set is representative of the distribution you're trying to learn. There are many tricks there one can use, like overweighting or replaying certain parts of the data. Or if you identify some gap in your data set, you could imagine that's where you put your synthetic generation capabilities to work on.

Dwarkesh Patel
Nowadays, people are paying attention to the RL stuff that DeepMind did many years before. What are the early research directions, or something that was done way back in the past, that you think will be a big deal but people just haven't been paying attention to it? There was a time where people weren't paying attention to scaling. What's the thing now that is totally underrated?

Demis Hassabis
Well, I think that the history of the last couple of decades has been things coming in and out of fashion, right? A while ago, maybe five-plus years ago, we were pioneering with AlphaGo and before that DQN. It was the first system that worked on Atari, our first big system really more than ten years ago now, that scaled up Q-learning and reinforcement learning techniques and combined that with deep learning to create deep reinforcement learning. We used that to scale up to master some pretty complex tasks like playing Atari games just from the pixels. I do actually think a lot of those ideas need to come back in again and, as we talked about earlier, combine them with the new advances in large models and large multimodal models, which are obviously very exciting as well. So I do think there's a lot of potential for combining some of those older ideas together with the newer ones.

Dwarkesh Patel
Is there any potential for the AGI to eventually come from a pure RL approach? The way we're talking about it, it sounds like the LLM will form the right prior and then this sort of tree search will go on top of that. Or is it a possibility that it comes completely out of the dark?

Demis Hassabis
Theoretically, I think there's no reason why you couldn't go full AlphaZero-like on it. There are some people here at Google DeepMind and in the RL community who work on that, fully assuming no priors, no data, and just building all knowledge from scratch. I think that's valuable because those ideas and those algorithms should also work when you have some knowledge too. Having said that, I think by far the quickest way to get to AGI, and the most plausible way, is to use all the knowledge that's existing in the world right now that we've collected from things like the Web. We have these scalable algorithms, like transformers, that are capable of ingesting all of that information. So I don't see why you wouldn't start with a model as a kind of prior, or to build on it and to make predictions that help bootstrap your learning. I just think it doesn't make sense not to make use of that. So my betting would be that the final AGI system will have these large multimodal models as part of the overall solution, but they probably won't be enough on their own. You'll need this additional planning search on top.

Dwarkesh Patel
This sounds like the answer to the question I'm about to ask. As somebody who's been in this field for a long time and seen different trends come and go, what do you think the strong version of the scaling hypothesis gets right and what does it get wrong? The idea that you just throw enough compute at a wide enough distribution of data and you get intelligence.

Demis Hassabis
My view is that this is kind of an empirical question right now. I think it was pretty surprising to almost everyone, including the people who first worked on the scaling hypotheses, how far it's gone. In a way, I look at the large models today and I think they're almost unreasonably effective for what they are. I think it's pretty surprising some of the properties that emerge. In my opinion, they've clearly got some form of concepts and abstractions and things like that. I think if we were talking five-plus years ago, I would have said to you that maybe we need an additional algorithmic breakthrough in order to do that, maybe more like how the brain works. I think that's still true if we want explicit abstract concepts, neat concepts, but it seems that these systems can implicitly learn that. Another really interesting, unexpected thing was that these systems have some sort of grounding even though they don't experience the world multimodally, at least until more recently when we have the multimodal models. The amount of information and models that can be built up just from language is surprising. I think that I'd have some hypotheses about why that is. I think we get some grounding through the RLHF feedback systems because obviously the human raters are by definition, grounded people. We're grounded in reality, so our feedback is also grounded. Perhaps there's some grounding coming in through there. Also if you're able to ingest all of it, maybe language contains more grounding than linguists thought before. So it actually raises some very interesting philosophical questions that people haven't even really scratched the surface of yet. Looking at the advances that have been made, it's quite interesting to think about where it's going to go next. In terms of your question of large models, I think we've got to push scaling as hard as we can and that's what we're doing here. It's an empirical question, whether that will hit an asymptote or a brick wall, and there are different people who argue about that. I think we should just test it. I think no one knows. In the meantime, we should also double down on innovation and invention. This is something where Google Research and DeepMind and Google Brain have pioneered many, many things over the last decade. That's our bread and butter. You can think of half our effort as having to do with scaling and half our efforts having to do with inventing the next architectures and the next algorithms that will be needed, knowing that larger and larger scaled models are coming down the line. So my betting right now, but it's a loose betting, is that you need both. I think you've got to push both of them as hard as possible and we're in a lucky position that we can do that.

Dwarkesh Patel
I want to ask more about the grounding. You can imagine two things that might change which would make the grounding more difficult. One is that as these models get smarter, they are going to be able to operate in domains where we just can't generate enough human labels, just because we're not smart enough. If it does a million-line pull request, how do we tell it, for example, this is within the constraints of our morality and the end goal we wanted and this isn't? The other thing has to do with what you were saying about compute. So far we've been doing next token prediction and in some sense it's a guardrail, because you have to talk as a human would talk and think as a human would think. Now, additional compute is

maybe going to come in the form of reinforcement learning where it's just getting to the objective and we can't really trace how you got there. When you combine those two, how worried are you that the grounding goes away?

Demis Hassabis
I think if it's not properly grounded, the system won't be able to achieve those goals properly. In a sense, you have to have some grounding for a system to actually achieve goals in the real world. I do actually think that these systems, and things like Gemini, are becoming more multimodal. As we start ingesting things like video and audiovisual data as well as text data, then the system starts correlating those things together. I think that is a form of proper grounding. So I do think our systems are going to start to understand the physics of the real world better.

Then one could imagine the active version of that as a very realistic simulation or game environment where you're starting to learn about what your actions do in the world and how that affects the world itself. The world stays itself, but it also affects what next learning episode you're getting. So these RL agents we've always been working on and pioneered, like AlphaZero and AlphaGo, actually are active learners. What they decide to do next affects what next learning piece of data or experience they're going to get. So there's this very interesting sort of feedback loop.

And of course, if we ever want to be good at things like robotics, we're going to have to understand how to act in the real world.

Dwarkesh Patel
So there's grounding in terms of whether the capabilities will be able to proceed, whether they will be enough in touch with reality to do the things we want. There's another sense of grounding in that we've gotten lucky that since they're trained on human thought, they maybe think like a human. To what extent does that stay true when more of the compute for training comes from just "did you get the right outcome" and it's not guardrailed by "are you proceeding on the next token as a human would?" Maybe the broader question I'll pose to you is, and this is what I asked Shane as well, what would it take to align a system that's smarter than a human? Maybe it thinks in alien concepts and you can't really monitor the million-line pull request because you can't really understand the whole thing and you can't give labels.

Demis Hassabis
This is something Shane and I, and many others here, have had at the forefront of our minds since before we started DeepMind because we planned for success. In 2010, no one was thinking about AI let alone AGI. But we already knew that if we could make progress with these systems and these ideas, the technology created would be unbelievably transformative. So we were already thinking 20 years ago about what the consequences of

that would be, both positive and negative. Of course, the positive direction is amazing science, things like AlphaFold, incredible breakthroughs in health and science, and mathematical and scientific discovery. But we also have to make sure these systems are sort of understandable and controllable.

This will be a whole discussion in itself, but there are many, many ideas that people have such as more stringent eval systems. I think we don't have good enough evaluations and benchmarks for things like if the system can deceive you. Can it exfiltrate its own code or do other undesirable behaviors? There are also ideas of using AI, not general learning ones but maybe narrow AIs that are specialized for a domain, to help us as the human scientists to analyze and summarize what the more general system is doing. So there's narrow AI tools. I think that there's a lot of promise in creating hardened sandboxes or simulations that are hardened with cybersecurity arrangements around the simulation, both to keep the AI in and to keep hackers out. You could experiment a lot more freely within that sandbox domain. There's many, many other ideas, including the analysis stuff we talked about earlier, where we can analyze and understand what the concepts are that this system is building and what the representations are. So maybe then they're not so alien to us and we can actually keep track of the kind of knowledge that it's building.

Dwarkesh Patel
Stepping back a bit, I'm curious what your timelines are. So Shane said his modal outcome is 2028. I think that's maybe his median. What is yours?

Demis Hassabis
I don't have prescribed specific numbers to it because I think there's so many unknowns and uncertainties. Human ingenuity and endeavor comes up with surprises all the time. So that could meaningfully move the timelines. I will say that when we started DeepMind back in 2010, we thought of it as a 20-year project. And I think we're on track actually, which is kind of amazing for 20-year projects because usually they're always 20 years away. That's the joke about whatever, quantum, AI, take your pick. But I think we're on track. So I wouldn't be surprised if we had AGI-like systems within the next decade.

Dwarkesh Patel
Do you buy the model that once you have an AGI, you have a system that basically speeds up further AI research? Maybe not in an overnight sense, but over the course of months and years you would have much faster progress than you would have otherwise had?

Demis Hassabis
I think that's potentially possible. I think it partly depends on what we, as a society, decide to use the first nascent AGI systems or proto-AGI systems for. Even the current LLMs seem to be pretty good at coding and we have systems like AlphaCode. We also have theorem proving systems. So one could imagine combining these ideas together and making them a

lot better. I could imagine these systems being quite good at designing and helping us build future versions of themselves, but we also have to think about the safety implications of that of course.

Dwarkesh Patel
I'm curious what you think about that. I'm not saying this is happening this year, but eventually you'll be developing a model where you think there's some chance that it'll be capable of an intelligence explosion-like dynamic once it's fully developed. What would have to be true of that model at that point where you're comfortable continuing the development of the system? Something like, "I've seen these specific evals, I've understood its internal thinking and its future thinking enough."

Demis Hassabis
We need a lot more understanding of the systems than we do today before I would even be confident of explaining to you what we'd need to tick box there. I think what we've got to do in the next few years, in the time before those systems start arriving, is come up with the right evaluations and metrics. Ideally formal proofs, but it's going to be hard for these types of systems, so at least empirical bounds around what these systems can do. That's why I think about things like deception as being quite root node traits that you don't want. If you're confident that your system is exposing what it actually thinks, then that opens up possibilities of using the system itself to explain aspects of itself to you. The way I think about that is like this. If I were to play a game of chess against Garry Kasparov, which I've played in the past, Magnus Carlsen, or the amazing chess players of all time, I wouldn't be able to come up with a move that they could. But they could explain to me why they came up with that move and I could understand it post hoc, right? That's the sort of thing one could imagine. One of the capabilities that we could make use of these systems is for them to explain it to us and even maybe get the proofs behind why they're thinking something, certainly in a mathematical problem.

Dwarkesh Patel
Got it. Do you have a sense of what the converse answer would be? So what would have to be true where tomorrow morning you're like "oh, man, I didn't anticipate this." You see some specific observation tomorrow morning that makes you say "we got to stop Gemini 2 training."

Demis Hassabis
I could imagine that. This is where things like the sandbox simulations are important. I would hope we're experimenting in a safe, secure environment when something very unexpected happens. There's a new unexpected capability or something that we didn't want. We explicitly told the system we didn't want it but then it did and it lied about it. These are the kinds of things where one would want to then dig in carefully. The systems that are around today are not dangerous, in my opinion, but in a few years they might have potential.

Then you would ideally pause and really get to the bottom of why it was doing those things before one continued.

Dwarkesh Patel
Going back to Gemini, I'm curious what the bottlenecks were in the development. Why not immediately make it one order of magnitude bigger if scaling works?

Demis Hassabis
First of all, there are practical limits. How much compute can you actually fit in one data center? You're also bumping up against very interesting distributed computing kind of challenges. Fortunately, we have some of the best people in the world working on those challenges and cross data center training, all of these kinds of things. There are very interesting hardware challenges and we have our TPUs that we're building and designing all the time as well as using GPUs. So there's all of that. Scaling laws also don't just work by magic. You still need to scale up the hyperparameters, and various innovations are going in all the time with each new scale. It's not just about repeating the same recipe at each new scale. You have to adjust the recipe and that's a bit of an art form. You have to sort of get new data points. If you try to extend your predictions and extrapolate them several orders of magnitude out, sometimes they don't hold anymore. There can be step functions in terms of new capabilities and some things hold, other things don't. Often you do need those intermediate data points to correct some of your hyperparameter optimization and other things, so that the scaling law continues to be true. So there are various practical limitations to that. One order of magnitude is probably about the maximum that you want to do between each era.

Dwarkesh Patel
That's so fascinating. In the GPT-4 technical report, they say that they were able to predict the training loss with a model with tens of thousands of times less compute than GPT-4. They could see the curve. But the point you're making is that the actual capabilities that loss implies may not be so.

Demis Hassabis
Yeah, the downstream capabilities sometimes don't follow. You can often predict the core metrics like training loss or something like that, but then it doesn't actually translate into MMLU, or math, or some other actual capability that you care about. They're not necessarily linear all the time. There are non-linear effects there.

Dwarkesh Patel
What was the biggest surprise to you during the development of Gemini in terms of something like this happening?

Demis Hassabis
I wouldn't say there was one big surprise. It was very interesting trying to train things at that size and learning about all sorts of things from an organizational standpoint, like how to babysit such a system and to track it. There's also things like getting a better understanding of the metrics you're optimizing versus the final capabilities that you want. I would say that's still not a perfectly understood mapping, but it's an interesting one that we're getting better and better at.

Dwarkesh Patel
There's a perception that maybe other labs are more compute-efficient than DeepMind has been with Gemini. I don't know what you make of that perception.

Demis Hassabis
I don't think that's the case. I think that actually Gemini 1 used roughly the same amount of compute, maybe slightly more, than what was rumored for GPT-4. I don't know exactly what was used but I think it was in the same ballpark. I think we're very efficient with our compute and we use our compute for many things. One is not just the scaling but, going back to earlier, more innovations and ideas. A new innovation, a new invention, is only useful if it can also scale. So you need quite a lot of compute to do new invention because you've got to test many things, at least some reasonable scale, and make sure that they work at that scale. Also, some new ideas may not work at a toy scale but do work at a larger scale. In fact, those are the more valuable ones. So if you think about that exploration process, you need quite a lot of compute to be able to do that. The good news is we're pretty lucky at Google. I think this year we're going to have the most compute by far of any sort of research lab. We hope to make very efficient and good use of that in terms of both scaling and the capability of our systems and also new inventions.

Dwarkesh Patel
What's been the biggest surprise to you, if you go back to yourself in 2010 when you were starting DeepMind, in terms of what AI progress has looked like? Did you anticipate back then that it would, in some large sense, amount to spending billions of dollars into these models? Or did you have a different sense of what it would look like?

Demis Hassabis
We thought that actually, and I know you've interviewed my colleague Shane. He always thought in terms of compute curves and comparing it roughly to the brain, how many neurons and synapses there are very loosely. Interestingly, we're actually in that kind of regime now with roughly the right order of magnitude of number of synapses in the brain and the sort of compute that we have. But I think more fundamentally, we always thought that we bet on generality and learning. So those were always at the core of any technique we would use. That's why we triangulated on reinforcement learning, and search, and deep learning as three types of algorithms that would scale, be very general, and not require a lot

of handcrafted human priors. We thought that was the sort of failure mode of the efforts to build AI in the 90s in places like MIT. There were very logic-based systems, expert systems, and masses of hand-coded, handcrafted human information going into them that turned out to be wrong or too rigid. So we wanted to move away from that and I think we spotted that trend early. Obviously, we used games as our proving ground and we did very well with that. I think all of that was very successful and maybe inspired others. AlphaGo, I think, was a big moment for inspiring many others to think "oh, actually, these systems are ready to scale." Of course then, with the advent of transformers, invented by our colleagues at Google Research and Brain, that was the type of deep learning that allowed us to ingest masses of amounts of information. That has really turbocharged where we are today. So I think that's all part of the same lineage. We couldn't have predicted every twist and turn there, but I think the general direction we were going in was the right one.

Dwarkesh Patel
It's fascinating if you read your old papers or Shane's old papers. In Shane's thesis in 2009, he said "well, the way we would test for AI is, can you compress Wikipedia?" And that's literally, the loss function for LLMs. Or in your own paper in 2016 before transformers, you were comparing neuroscience and AI and you said attention is what is needed.

Demis Hassabis
Exactly. So we had these things called out and we had some early attention papers, but they weren't as elegant as transformers in the end, neural Turing machines and things like this. Transformers were the nicer and more general architecture of that.

Dwarkesh Patel
When you extrapolate all this out forward and you think about superhuman intelligence, what does that landscape look like to you? Is it still controlled by a private company? What should the governance of that look like concretely?

Demis Hassabis
I think that this is so consequential, this technology. I think it's much bigger than any one company or even industry in general. I think it has to be a big collaboration with many stakeholders from civil society, academia, government, etc. The good news is that with the popularity of the recent chatbot systems, I think that has woken up many of these other parts of society to the fact that this is coming and what it will be like to interact with these systems. And that's great. It's opened up lots of doors for very good conversations. An example of that was the safety summit the UK hosted a few months ago, which I thought was a big success in getting this international dialogue going. I think the whole of society needs to be involved in deciding what we want to deploy these models for? How do we want to use them and what do we not want to use them for? I think we've got to try and get some international consensus around that and also make sure that these systems benefit everyone, for the good of society in general. That's why I push so hard for things like AI for

science. I hope that with things like our spin-out, Isomorphic, we're going to start curing terrible diseases with AI, accelerate drug discovery, tackle climate change, and do other amazing things. There are big challenges that face humanity, massive challenges. I'm actually optimistic we can solve them because we've got this incredibly powerful tool of AI coming down the line that we can apply to help us solve many of these problems. Ideally, we would have a big consensus around that and a big discussion at sort of the UN level if possible.

Dwarkesh Patel
One interesting thing is if you look at these systems and chat with them, they're immensely powerful and intelligent. But it's interesting the extent to which they haven't automated large sections of the economy yet. Whereas if five years ago I showed you Gemini, you'd be like "wow, this is totally coming for a lot of things." So how do you account for that? What's going on that it hasn't had the broader impact yet?

Demis Hassabis
I think that just shows we're still at the beginning of this new era. I think there are some interesting use cases where you can use these chatbot systems to summarize stuff for you and do some simple writing, maybe more boilerplate-type writing. But that's only a small part of what we all do every day. I think for more general use cases we still need new capabilities, things like planning and search but also things like personalization and episodic memory. That's not just long context windows, but actually remembering what we spoke about 100 conversations ago. I'm really looking forward to things like recommendation systems that help me find better, more enriching material, whether that's books or films or music and so on. I would use that type of system every day. So I think we're just scratching the surface of what these AI assistants could actually do for us in our general, everyday lives and also in our work context as well. I think they're not reliable yet enough to do things like science with them. But I think one day, once we fix factuality and grounding and other things, I think they could end up becoming the world's best research assistant for you as a scientist or as a clinician.

Dwarkesh Patel
I want to ask about memory. You had this fascinating paper in 2007 where you talked about the links between memory and imagination and how they, in some sense, are very similar. People often claim that these models are just memorizing. How do you think about that claim? Is memorization all you need because in some deep sense, that's compression? What's your intuition here?

Demis Hassabis
At the limit, one maybe could try and memorize everything but it wouldn't generalize out of your distribution. The early criticisms of these early systems were that they were just regurgitating and memorizing. I think clearly in the Gemini, GPT-4 type era, they are

definitely generalizing to new constructs. Actually my thesis, and that paper particularly that started that area of imagination in neuroscience, was showing that first of all memory, at least human memory, is a reconstructive process. It's not a videotape. We sort of put it together back from components that seem familiar to us, the ensemble. That's what made me think that imagination might be the same thing. Except in this case you're using the same semantic components, but now you're putting it together in a way that your brain thinks is novel, for a particular purpose like planning. I do think that that kind of idea is still probably missing from our current systems, pulling together different parts of your world model to simulate something new that then helps with your planning, which is what I would call imagination.

Dwarkesh Patel
For sure. Now you guys have the best models in the world with the Gemini models. Do you plan on putting out some sort of framework like the other two major AI labs have? Something like "once we see these specific capabilities, unless we have these specific safeguards, we're not going to continue development or we're not going to ship the product out."

Demis Hassabis
Yes, we already have lots of internal checks and balances but we're going to start publishing. Actually, watch this space. We're working on a whole bunch of blog posts and technical papers that we'll be putting out in the next few months along similar lines of things like responsible scaling laws and so on. We have those implicitly internally in various safety councils that people like Shane chair and so on. But it's time for us to talk about that more publicly I think. So we'll be doing that throughout the course of the year.

Dwarkesh Patel
That's great to hear. Another thing I'm curious about is, there's not only the risk of the deployed model being something that people can use to do bad things, but there's also rogue actors, foreign agents, and so forth, being able to steal the weights and then fine-tune them to do crazy things. How do you think about securing the weights to make sure something like this doesn't happen, making sure a very key group of people has access to them?

Demis Hassabis
It's interesting. First of all, there's two parts. One is security, one is open source, which maybe we can discuss. The security is super key just as normal cybersecurity type things. I think we're lucky at Google DeepMind. We're behind Google's firewall and cloud protection which I think is best in class in the world corporately. So we already have that protection. Behind that, we have specific DeepMind protections within our code base. It's sort of a double layer of protection. So I feel pretty good about that. You can never be complacent on that but I feel it's already the best in the world in terms of cyber defenses. We've got to carry

on improving that and again, things like the hardened sandboxes could be a way of doing that as well. Maybe there are even specifically secure data centers or hardware solutions to this too that we're thinking about. I think that maybe in the next three, four, five years, we would also want air gaps and various other things that are known in the security community. So I think that's key and I think all frontier labs should be doing that because otherwise for rogue nation-states and other dangerous actors, there would obviously be a lot of incentive for them to steal things like the weights. Of course, open source is another interesting question. We're huge proponents of open source and open science. We've published thousands of papers, things like AlphaFold and transformers and AlphaGo. All of these things we put out there into the world, published and open source, most recently GraphCast, our weather prediction system. But when it comes to the general-purpose foundational technology, I think the question I would have for open source proponents is, how does one stop bad actors, individuals or up to rogue states, taking those same open source systems and repurposing them for harmful ends? We have to answer that question. I don't know what the answer is to that, but I haven't heard a compelling, clear answer to that from proponents of just open sourcing everything. So I think there has to be some balance there. Obviously, it's a complex question of what that is.

Dwarkesh Patel
I feel like tech doesn't get the credit it deserves for funding hundreds of billions of dollars' worth of R&D, obviously you have DeepMind with systems like AlphaFold and so on. When we talk about securing the weights, as we said maybe right now it's not something that is going to cause the end of the world or anything, but as these systems get better and better, there's the worry that a foreign agent or something gets access to them. Presumably right now there's dozens to hundreds of researchers who have access to the weights. What's a plan for getting the weights in a situation room where if you need to access them it's some extremely strenuous process and no individual can really take them out?

Demis Hassabis
One has to balance that with allowing for collaboration and speed of progress. Another interesting thing is that of course you want brilliant independent researchers from academia or things like the UK AI Safety Institute and the US one to be able to red team these systems. So one has to expose them to a certain extent, although that's not necessarily the weights. We have a lot of processes in place about making sure that only if you need them, those people who need access have access. Right now, I think we're still in the early days of those kinds of systems being at risk. As these systems become more powerful and more general and more capable, I think one has to look at the access question.

Dwarkesh Patel
Some of these other labs have specialized in different things relative to safety, Anthropic for example with interpretability. Do you have some sense of where you guys might have an

edge? Now that you have the frontier model, where are you guys going to be able to put out the best frontier research on safety?

Demis Hassabis
I think we helped pioneer RLHF and other things like that which can obviously be used for performance but also for safety. I think that a lot of the self-play ideas and these kinds of things could also be used to auto-test a lot of the boundary conditions that you have with the new systems. Part of the issue is that with these very general systems, there's so much surface area to cover about how these systems behave. So I think we are going to need some automated testing. Again, with things like simulations and games, very realistic virtual environments, I think we have a long history of using those kinds of systems and making use of them for building AI algorithms. I think we can leverage all of that history. And then around Google, we're very lucky to have some of the world's best cybersecurity experts, hardware designers. I think we can bring that to bear for security and safety as well.

Dwarkesh Patel
Let's talk about Gemini. So now you guys have the best model in the world. I'm curious. The default way to interact with these systems has been through chat so far. Now that we have multimodal and all these new capabilities, how do you anticipate that changing? Do you think that'll still be the case?

Demis Hassabis
I think we're just at the beginning of actually understanding how exciting that might be to interact with a full multimodal model system. It'll be quite different from what we're used to today with the chatbots. I think the next versions of this over the next year, 18 months, we'll maybe have some contextual understanding of the environment around you through a camera or a phone or some glasses. I could imagine that as the next step. And then I think we'll start becoming more fluid in understanding "let's sample from a video, let's use voice." Maybe even eventually things like touch and if you think about robotics, other types of sensors. So I think the world's about to become very exciting in the next few years as we start getting used to the idea of what true multimodality means.

Dwarkesh Patel
On the robotics subject, when he was on the podcast Ilya said that the reason OpenAI gave up on robotics was because they didn't have enough data in that domain, at least at the time they were pursuing it. You guys have put out different things like Robo-Transformer and other things. Do you think that's still a bottleneck for robotics progress, or will we see progress in the world of atoms as well as the world of bits?

Demis Hassabis
We're very excited about our progress with things like Gato and RT-2. We've always liked robotics and we've had amazing research in that. We still have that going now because we like the fact that it's a data-poor regime. That pushes us in very interesting research directions that we think are going to be useful anyway: sampling efficiency and data efficiency in general, transfer learning, learning from simulation and transferring that to reality, sim-to-real. All of these are very interesting general challenges that we would like to solve. The control problem. So, we've always pushed hard on that. I think Ilya is right. It is more challenging because of the data problem. But I think we're starting to see the beginnings of these large models being transferable to the robotics regime. They can learn in the general domain, language domain and other things, and then just treat tokens like Gato as any type of token. The token could be an action, it could be a word, it could be part of an image, a pixel, or whatever it is. That's what I think true multimodality is. To begin with, it's harder to train a system like that than a straightforward language system. But going back to our early conversation on transfer learning, you start seeing that with a true multimodal system, the other modalities benefit some different modalities. You get better at language because you now understand a little bit about video. So I do think it's harder to get going, but ultimately we'll have a more general, more capable system like that.

Dwarkesh Patel
What ever happened to Gato? That was super fascinating that you could have it play games and also do video and also do text.

Demis Hassabis
We're still working on those kinds of systems, but you can imagine we're trying to build those ideas into our future generations of Gemini to be able to do all of those things. Robotics, transformers, and things like that, you can think of them as follow-ups to that.

Dwarkesh Patel
Will we see asymmetric progress in the domains in which the self-play kinds of things you're talking about will be especially powerful? So math and code. Recently, you have these papers out about this. You can use these things to do really cool, novel things. Will they be superhuman coders, but in other ways they might still be worse than humans? How do you think about that?

Demis Hassabis
I think that we're making great progress with math and things like theorem proving and coding. But it's still interesting if one looks at creativity in general, and scientific endeavor in general. I think we're getting to the stage where our systems could help the best human scientists make their breakthroughs quicker, almost triage the search space in some ways. Perhaps find a solution like AlphaFold does with a protein structure. They're not at the level where they can create the hypothesis themselves or ask the right question. As any top

scientist will tell you, the hardest part of science is actually asking the right question. It's boiling down that space to the critical question we should go after and then formulating the problem in the right way to attack it. That's not something our systems really have any idea how to do, but they are suitable for searching large combinatorial spaces if one can specify the problem with a clear objective function. So that's very useful already for many of the problems we deal with today, but not the most high-level creative problems.

Dwarkesh Patel
DeepMind has published all kinds of interesting stuff in speeding up science in different areas. If you think AGI is going to happen in the next 10 to 20 years, why not just wait for the AGI to do it for you? Why build these domain-specific solutions?

Demis Hassabis
I think we don't know how long AGI is going to be. We always used to say, back even when we started DeepMind, that we don't have to wait for AGI in order to bring incredible benefits to the world. My personal passion especially has been AI for science and health. You can see that with things like AlphaFold and all of our various Nature papers on different domains and material science work and so on. I think there's lots of exciting directions and also impact in the world through products too. I think it's very exciting and a huge unique opportunity we have as part of Google. They've got dozens of billion-user products that we can immediately ship our advances into and then billions of people can improve, enrich, and enhance their daily lives. I think it's a fantastic opportunity for impact on all those fronts.
I think the other reason from the point of view of AGI specifically is that it battle tests your ideas. You don't want to be in a research bunker where you theoretically are pushing things forward, but then actually your internal metrics start deviating from real-world things that people would care about, or real-world impact. So you get a lot of direct feedback from these real-world applications that then tells you whether your systems really are scaling or if we need to be more data efficient or sample efficient. Because most real-world challenges require that. So it kind of keeps you honest and pushes you to keep nudging and steering your research directions to make sure they're on the right path. So I think it's fantastic. Of course, the world benefits from that. Society benefits from that on the way, maybe many years before AGI arrives.

Dwarkesh Patel
The development of Gemini is super interesting because it comes right at the heels of merging these different organizations, Brain and DeepMind. I'm curious, what have been the challenges there? What have been the synergies? It's been successful in the sense that you have the best model in the world now. What's that been like?

Demis Hassabis
It's been fantastic actually, over the last year. Of course it's been challenging to do, like any big integration coming together. You're talking about two world-class organizations with

long, storied histories of inventing many important things from deep reinforcement learning to transformers. So it's very exciting to actually pool all of that together and collaborate much more closely. We always used to be collaborating, but more on a project-by-project basis versus a much deeper, broader collaboration like we have now. Gemini is the first fruit of that collaboration, including the name Gemini implying twins. Of course, a lot of other things are made more efficient like pooling compute resources together and ideas and engineering. I think at the stage we're at now, there are huge amounts of world-class engineering that have to go into building the frontier systems. I think it makes sense to coordinate that more.

Dwarkesh Patel
You and Shane started DeepMind partly because you were concerned about safety. You saw AGI coming as a live possibility. Do you think the people who were formerly part of Brain, that half of Google DeepMind now, approach it in the same way? Have there been cultural differences there in terms of that question?

Demis Hassabis
This is one of the reasons we joined forces with Google back in 2014. I think the entirety of Google and Alphabet, not just Brain and DeepMind, takes these questions of responsibility very seriously. Our kind of mantra is to try and be bold and responsible with these systems. I'm obviously a huge techno-optimist but I want us to be cautious given the transformative power of what we're bringing into the world collectively. I think it's important. It's going to be one of the most important technologies humanity will ever invent. So we've got to put all our efforts into getting this right and be thoughtful and also humble about what we know and don't know about the systems that are coming and the uncertainties around that. In my view, the only sensible approach when you have huge uncertainty is to be cautiously optimistic and use the scientific method to try and have as much foresight and understanding about what's coming down the line and the consequences of that before it happens. You don't want to be live A/B testing out in the world with these very consequential systems because unintended consequences may be quite severe. So I want us to move away, as a field, from a sort of "move fast and break things attitude" which has maybe served the Valley very well in the past and obviously created important innovations. I think in this case we want to be bold with the positive things that it can do and make sure we advance things like medicine and science whilst being as responsible and thoughtful as possible with mitigating the risks.

Dwarkesh Patel
That's why it seems like the responsible scaling policies are something that are a very good empirical way to pre-commit to these kinds of things.

Demis Hassabis
Yes, exactly.

Dwarkesh Patel
When you're doing these evaluations and for example it turns out your next model could help a layperson build a pandemic-class bioweapon or something, how would you think first of all about making sure those weights are secure so that they don't get out? And second, what would have to be true for you to be comfortable deploying that system? How would you make sure that this latent capability isn't exposed?

Demis Hassabis
The secure model part I think we've covered with the cybersecurity and making sure that's world-class and you're monitoring all those things. I think if a capability like that was discovered through red teaming or external testing, independent testers like government institutes or academia or whatever, then we would have to fix that loophole. Depending on what it was, that might require a different kind of constitution perhaps, or different guardrails, or more RLHF to avoid that. Or you could remove some training data, depending on what the problem is. I think there could be a number of mitigations. The first part is making sure you detect it ahead of time. So that's about the right evaluations and right benchmarking and right testing. Then the question is how one would fix that before you deployed it. But I think it would need to be fixed before it was deployed generally, for sure, if that was an exposure surface.

Dwarkesh Patel
Final question. You've been thinking in terms of the end goal of AGI at a time when other people thought it was ridiculous in 2010. Now that we're seeing this slow takeoff where we're actually seeing generalization and intelligence, what is like psychologically seeing this? What has that been like? Has it just been sort of priced into your world model so it's not new news for you? Or actually just seeing it live, are you like "wow, something's really changed"? What does it feel like?

Demis Hassabis
For me, yes, it's already priced into my world model of how things were going to go, at least from the technology side. But obviously, we didn't necessarily anticipate that the general public would be so interested this early in the sequence. If ChatGPT and chatbots hadn't gotten the interest they ended up getting—which I think was quite surprising to everyone that people were ready to use these things even though they were lacking in certain directions, impressive though they are—then we would have produced more specialized systems built off of the main track, like AlphaFold and AlphaGo, our scientific work. I think then the general public maybe would have only paid attention later down the road when in a few years' time, we have more generally useful assistant-type systems. So that's been interesting. That's created a different type of environment that we're now all operating in as a field. It's a little bit more chaotic because there's so many more things going on, and there's so much VC money going into it, and everyone's sort of almost losing their minds over it. The only thing I worry about is that I want to make sure that, as a field, we act

responsibly and thoughtfully and scientifically about this and use the scientific method to approach this in an optimistic but careful way. I think I've always believed that that's the right approach for something like AI, and I just hope that doesn't get lost in this huge rush.

Dwarkesh Patel
Well, I think that's a great place to close. Demis, thank you so much for your time and for coming on the podcast.

Demis Hassabis
Thanks. It's been a real pleasure.