**CRAIG:** Hi, I’m Craig Smith and this is Eye on AI.

Before we begin, I want to introduce you to [ClearML](https://t.clear.ml/eye_on_ai), an open-source MLOps solution. They’ll be sponsoring the podcast for a while and I really appreciate their support. Do me a favor and give them a try. Users love to customize the solution, which allows you to easily Track, Orchestrate, and Automate ML Workflows at Scale.

Whether you're a Data Engineer, ML engineer, DevOps, or a Data Scientist, [ClearML](https://t.clear.ml/eye_on_ai) is hands-down the best collaborative MLOps tool with full visibility and extensibility.

Setup your own server or Signup for free, log your entire process, version data, build model repositories, provision machines, schedule containers, and deploy pipelines directly from code.

Supercharge your entire development process - Go checkout ClearML today – at [clear.ml](https://t.clear.ml/eye_on_ai)!

This week, I talk to Pushmeet Kohli, the head of DeepMind’s AI for Science and one of the brains behind AlphaFold, the machine learning system that is helping solve the protein folding problem. Pushmeet talked about how AlphaFold developed, where it is headed and about some of the other amazing things his team is doing at DeepMind.

I hope you find the conversation as exciting as I did.

**CRAIG:** So, Pushmeet, can you start by introducing yourself and how you got to DeepMind and what you do at DeepMind?

**PUSHMEET:** Yeah, thanks Craig for having me.

**PUSHMEET:** My name is Pushmeet Kohli. I lead the science program at DeepMind. I joined DeepMind roughly five years ago now. My background is in computer science. I did my undergrad in computer science and then my research work and thesis work, PhD work was focused on discreet optimization and applying it to specific problems in machine learning.

**PUSHMEET:** And over the last 15, 20 years I have worked on machine learning and used it a variety of different application areas from information retrieval, to computer vision to software analysis and testing. And for the last five years, I've been very excited to apply to problems that we encounter in the natural sciences and other applications that are important.

**CRAIG:** It seems that your history is very much on the application side. How did you end up on the more basic research side.

**PUSHMEET:** My approach to research has been to consider a problem that is of interest. And that is important and to think about what makes that problem challenging and how can you approach it in a multi-disciplinary way. When I moved to DeepMind, I wanted to have that problem driven approach to the work I did.

**PUSHMEET:** Demis, the founder of DeepMind, and Koray (Kavukcuoglu), who's the VP of research, asked me to consider AI for science as the next thing that I take on and it came to me that it is very much like the applied approach that I was taking, because here the problems are in the natural sciences where we are taking a multi disciplinary approach and a scientific approach towards technically very hard challenges through leveraging developments in machine learning and AI. So it's still AI applied towards a particular domain, but now in the service of science, rather than an application, like information retrieval or computer vision.

**CRAIG:** On AlphaFold, can you take us back? What happened between AlphaFold 1 and AlphaFold 2. And what you're looking at for, presumably, AlphaFold 3? Because there obviously was a huge leap between CASP13 and CASP14.

**PUSHMEET:** So CASP is the particular assessment for all protein structures. It has been going on for a number of years. Like the CASP 13 is the 13 th edition of that assessment. And it was the first time we actually participated in this assessment.

**PUSHMEET:** So the protein structure prediction problem is given a protein, which is essentially a sequence of amino acids, right? So these amino acids, you have these 20 amino acids, which are like the letters of the alphabet and there is a particular sequence. And you're given the sequence and you are asked, what is the structure of the whole molecule in 3D. And that structure in many cases, informs the function of that protein.

**PUSHMEET:** Given what has happened in the last couple of years, the general public is quite familiar with the spike protein on Sars-CoV-2. That spike protein is basically a protein, which has a specific structure, which then binds to the ACE 2 receptor and so on. So understanding the structure is very important from a biological perspective, to understand the function of the protein.

**PUSHMEET:** Now in CASP 13, we took the approach that we would trained a neural network, which given a sequence of amino acids are given a protein sequence. It will try to predict what is the distance between any two amino acids.

**PUSHMEET:** Once we have the distance, then we have a second component, which will try to infer the structure of the protein, given these distance constraints that we have.

**PUSHMEET:** And so this was a two module process. The first module employed a neural network, which was reading in a sequence

**PUSHMEET:** and some evolutionary history associated with the sequence to make predictions about the distance between different amino acid residues. And then there was a second module, which was taking those distance constraints to find the overall 3D structure. And that performed quite well compared to what was the state of the art .

**PUSHMEET:** We were best in the world, but we were still quite far from actually being useful and being close to the accuracy level that an experimental approach would give you.

**PUSHMEET:** So we approached CASP 14 by taking a very risky approach. At that time, basically it looked very risky, but now in hindsight, it was the right thing to do, which was to go back to the machine learning philosophy and make sure that the probability distributions, the errors propagate end to end and learning happens across the whole system.

**PUSHMEET:** We really want to go from the sequence to the structure. And if there are any errors in the structure, then we want to backtrack and be able to associate In the model what was responsible for those errors and correct them.

**PUSHMEET:** This whole idea of backpropagation and do that learning across this whole spectrum. And so that was a very, a very big departure from the system which we had from AlphaFold 1. And when we started working on that system, Its level of performance was quite below what we had at CASP 13. And there were many months where it remained below and we were making steady progress, but it was quite hard work, but the underlying philosophy that this is a better approach to take and this would eventually win kept us going on that exploration path. And the result are now there to see, that actually this was the right architecture. And eventually this led us to the level of performance now that some people mentioned that it's like solving the protein folding problem.

**CRAIG:** Actually, the press that came out of that was a little confusing to me because that is the language that was used, that the protein folding problem has been solved. And certainly AlphaGo's results were far above those of other teams, But it's still a long way from being able to predict protein structures from an amino acid sequence and being able to design new proteins from new sequences. Is that where you're going? How much better do the results have to be, to get to that point and at the current level, when you say that it's useful, in what way is it useful?

**PUSHMEET:** Okay. So if you look at the number of possible proteins that can exist in the universe, that's basically infinite because you have the 20 amino acids, and you can arrange them in any sort of order, and you can make the protein as large as possible.

**PUSHMEET:** So there are exponential number of these proteins. Now, the second question is what does AlphFold predict well on? So in the critical assessment, CASP 14, the proteins that are selected by the assessors, these independent assessors are proteins that are extremely relevant.

**PUSHMEET:** For some biological application. So researchers have spent a lot of time and effort in trying to figure out the structure of these proteins using structured determination methods. Like , whether it's x-ray crystallography and then submitted these predictions confidentially to the organizers, without anyone knowing that these were the predictions.

**PUSHMEET:** And AlphaFold was first on this distribution of proteins. And on this distribution of proteins, it showed that its level of performance was very high. Like almost like close to a 90 GTT, a level, which now allows you to really carefully inform what could be the biological mechanisms in that.

**PUSHMEET:** Now on completely other distribution for so on distribution of proteins, which are not naturally, or evolutionary or related to natural sort of proteins, they are arbitrary proteins. They are the performance of AlphaFold is different and that's not what AlphaFold was trained on. So this is, where AlphaFold needs to spend more time.

**PUSHMEET:** And we see that for some of these proteins it performed really well, but on some, it doesn't. So on that distribution of proteins, which are, say new proteins, which have never been seen in nature, there we have to just do more work in understanding what would be the structure of those proteins.

**PUSHMEET:** So that's one aspect. What can ALphFold do today and what it needs more work on. Now, the second question is how do scientists use these structures? And it was actually a very surprising thing for us because we thought that we'd be working on AlphaFold and later on, we will see our biologists, experimentalists leverage this tool over the years, but even during the critical assessment during the assessment process itself AlphaFold was able to make new scientific contributions. Some experimental assessors had submitted a sequence of a protein in the hope that they would find the structure of the protein in time, for when the assessment happened. But it just turned out that they had the experimental data, but it was very hard to find the structure of the protein. So there is a thing called molecular replacement, where if you have an initial guess, then you can use that initial guess to refine the actual ground truth structure.

**PUSHMEET:** So when they looked at the AlphaFold prediction, they could use that AlphaFold prediction to explain their experimental data and thereby find the ground truth. In that way for proteins, which had completely unknown structure AlphaFold was able to provide that structure during the assessment process.

**PUSHMEET:** And since then, when SarsCov2 was first sequenced in early last year around February, we actually ran all the proteins of SarsCov2 through AlphaFold and looked at proteins for which no structural information was available.

**PUSHMEET:** So of course the spike had been imaged using varying techinques, but only certain elements of the spike, but many other proteins had no structural information associate with them. And so we made structural predictions using AlphaFold for those proteins. And eventually later on when researchers could see months later that the predictions that ALphaFold had made actually conformed. We are already seeing a lot of different applications of AlphaFold. there are a couple of partnerships which I can talk to you about. One is with the center for enzyme innovation, and they are looking at developing enzymes for decomposing plastics and their knowledge of the structure of the protein actually helps them in developing better enzymes for plastic decomposition. And then there is also partnerships we are doing with DNDi (Drugs for Neglected Diseases Initiative), which is looking at neglected diseases. And these are diseases which for a variety of reasons, have not attracted a lot of scientific research. And they happen in developing countries where there are less financial incentives to commercial entities to invest in fundamental research. And by providing the structural information, the idea is that we will accelerate the development of drugs or treatments and better understanding of diseases like Chagas and leishmaniasis.

**CRAIG:** That's interesting. In one of the blog posts you talk about in five years time, you expect to predict structures that would correspond to the gross topology problem, which I presume is just a more accurate depiction of the protein shape or fold. Is that where you're going with AlphaFold 3? 3 what AlphaFold 2 does it basically produces the structure of the protein. But these proteins, they are like Lego pieces in some way, because they're building blocks, which construct everything. Like they are basic building blocks of life. But also they are not Lego pieces in a too rigid sense. They can be flexible. They can articulate. There are multiple confirmations when they bind with a ligand or a small molecule, they might readjust their structure. When they come together with other proteins, they might adjust their structure. So they are like flexible Lego blocks and which can also change their shape. So there's a lot of research that can go into multiple confirmations, the different shapes that proteins reside in.

**PUSHMEET:** They can be research in the actual sort of dynamics of how proteins interact. And then they could be research on how two proteins come together and form complexes, because in all in big living systems, proteins are not working in isolation. They're working as part of a complex to build bigger machines, which are then accomplishing some function in the, in the body.

**PUSHMEET:** So there's a lot of research that still needs to be done in terms of the behavior and the structure and the mechanisms at play all from going from the smaller protein building blocks to the cellular and the system level.

**CRAIG:** Yeah. But will there be an Alpha 3 in the CASP 15, will it be alpha 2 with some refinements?

**PUSHMEET:** Yeah. So this is a matter of research Craig. What is the structure that will be required or what is the model that will be required for taking on these other challenges is a question of active research.

MUSIC

**CRAIG:** On the glassy dynamics problem or the model that you've built to understand glassy dynamics - and I'll let you define for listeners what the glassy dynamic problem is - I've read the paper in the blog post. I just wanted to ask one thing that fascinated me. is that you You talked about examining the models to understand what they've learned about the glassy dynamics system, a system with glassy dynamics. And I was wondering how do you do that? How do you examine a model to understand what it has learned?

**CRAIG:** Can you walk us through that.

**PUSHMEET:** Yeah. So let me first talk about the glassy dynamics project. Many systems, many sort of materials that you see around the world, around you. Sometimes they can be of a form of a crystal. So a crystal, it gives a sort of a regular grid. There are particular points.

**PUSHMEET:** And certain items are placed on those points. And that's structured is rigid. These sort of atoms are not moving around and shifting places or like just floating around in that infrastructure. They are, they have this fixed rigid places where they need to be positioned. In a glass that is not the case in a glass, basically that you have these molecules. They are now tightly coupled, but they don't have their fixed, assigned position. They are moving, but very slowly. And the key question is they're moving very slowly, but how are they going to move? Or what is, what can you predict what type of movement will happen? Because if you're looking at a glass and glasses are now everywhere around us from cell phones to laptop screens to cameras. They are used in construction, the many different types of glasses that we use all around us. And yet these are, they are not stuck in a certain shape. They are moving, they're moving at a very slow sort of time in a very slow way, but they are actually moving objects internally.

**PUSHMEET:** Their structure is changing. So understanding how the structure is changing is the fundamental scientific challenge for the problem of glassy systems. And so that was the task we gave a machine learning and AI system to reason about that. Given a structure, can you make predictions about how a different molecule will move in that structure. And we showed that a machine learning system, which was based on a graph neural network, was able to infer what would be and was able to infer these connections and make predictions about how things would move. So that was the glassy dynamics project.

**PUSHMEET:** But the other question that you asked me was what did we learn from it? And how do we analyse what the model has tried to do. And this also brings to some of our more recent work where we have shown how AI and machine learning methods have allowed the top mathematicians to discover surprising new results in fundamental math.

**PUSHMEET:** And the key question is science is not just about making predictions. It's also about being able to communicate how the world works, to develop theories about how the world operates communication and the ability to explain the mechanisms at play become an intrinsic part of the act of doing science.

**PUSHMEET:** So in the context of glassy dynamics, and this is is much more true for the AI formats work is we then look into the model and try to see what elements did it rely on to make predictions and what were the things or the concepts it had learned to be able to make those predictions and then those give insights to us.

**PUSHMEET:** And in the case of the maths projects to the mathematicians as to what might be happening and what might be the key results, discovered relationships, that existed in that area.

**CRAIG:** And my question is how do you do that? A neural network is an array of virtual nodes and edges. With weights and biases.

**CRAIG:** Do you have a way of scanning to see what weights or biases have changed since the last time you ran data through the network? Or is there a tool that you use to, to visualize that?

**PUSHMEET:** Yeah, so there are a variety of techniques that we use for this, but the fundamental sort of abstraction that you should think about is these techniques are trying to think about attribution.

**PUSHMEET:** So they are trying to see in making this prediction, what was important, what elements of the input were important in making this prediction? And if I change the inputs in this way, how would the prediction change. And by understanding those connections, that actually, this is how the inputs and outputs are related to each other in a more simplified sort of sense, we are able to discover new things.

**PUSHMEET:** For example, in the glassy dynamics approach, we were looking at the structure of the molecules, how they are arranged and how did that structural neighborhood structure allow us to make predictions about how the molecule would move. In the maths project, we were looking at, there are certain quantities that become extremely important of this mathematical object in our ability to predict these other property of this mathematical object.

**PUSHMEET:** So maybe there is some deep connection between these two sets of properties, which nobody knew about.

**PUSHMEET:** So, in some sense, by looking at the sensitivity and by isolating where the attention was of the model, by following the attention of the model, we are able to - I mean attention in a abstraction. There's also a more sort of machine learning version of attention. So I'm using attention in the abstract - but following the attention of the model, we are able to then think about what are the quantities that might be related.

**CRAIG:** How do you disseminate this work? Certainly. The machine learning community follows what you do, but for example, with the math community or with the biochemistry community, how long does it take for these kinds of developments to permeate and really change the the landscape in those communities.

**PUSHMEET:** Yeah. So our experience has been very positive in this regard. In any of the projects that we have done, especially in the space of science, the scientific process forces you to keep an open mind about different hypotheses at play. And so when you get more information about a particular approach or a particular direction in tackling the problem that is promising, then you are able to quickly reevaluate your strategy forward.

**PUSHMEET:** So our experience has been that the take-up of these ideas of machine learning and AI across the board, in all the scientific disciplines that we have currently worked in, has been very high. People are really keen to leverage these ideas in pursuit of the problems that they are working on.

**PUSHMEET:** And so from our end, we have taken a very multidisciplinary approach to this. And and the idea has been that this is not machine learning for X. Every project requires collaboration between the domain experts and the machine learning experts and coming together and solving the problem together rather than just taking some data and throwing machine learning or AI techniques at it.

**CRAIG:** And at the practical level do all of these organizations now have adequate machine learning teams to take the lessons from DeepMind and apply them to their domain because certainly a chemist is not a computer scientist. And DeepMind, I don't think is in the business of consulting. So each of these domains, I would guess are adopting machine learning and machine learning teams to aid in their research.

**PUSHMEET:** Yeah, absolutely. I think basically there are computational chemists already. Computation has been there for a while, right, in in all these different disciplines. So a lot of disciplines are now taking machine learning extremely seriously and are investing a lot.

**PUSHMEET:** In all the projects that we are working on, we have made our models and our technique available for the community. This is especially true for the structures of the whole human proteome and various other sort of organisms. We have already shared with the community as well as made a commitment to make structures available for a much larger universe of proteins.

**PUSHMEET:** And we are seeing the take-up of these techniques across the board. Whether it's in structural biology or computational chemistry or glassy dynamics. Researchers in all these disciplines are really starting to leverage machine learning and AI techniques in a very deep and sophisticated way.

**CRAIG:** How do you decide what problems to go after?

**PUSHMEET:** In DeepMind, we have a very project driven structure. We want the ability to really accelerate projects if they are promising. So the team has its own engineering and research engineering elements. And then there are top teams focused in particular domains.

**PUSHMEET:** So we have a protein structural biology and protein dynamics team. We have a team working on Quantum chemistry. We have a, sub-team looking at genomics. We have a, sub-team looking at math and so on.

**PUSHMEET:** So each of the scientific areas we are interested in, we have experts who are looking in those domains. And then we have a centralized engineering and research engineering team, which is looking at how we can take some of the ideas of machine learning in pursuit of those problems

**CRAIG:** and is there a team that nominates problems to you ,is there then like a committee that decides whether to commit resources to it because obviously there's endless problems that need attention.

**PUSHMEET:** At DeepMind, our fundamental research philosophy is taking on extremely hard and challenging problems. The way I think about it is in the scientific ecosystem, what is our role? If there is a PhD student working in a university on a particular problem and has a fair chance of cracking that problem, then that is not the right problem for us.

**PUSHMEET:** We are uniquely positioned to take on problems that require a lot of effort, which require multidisciplinary research, which requires a large team to work together, to take on this extremely large and impactful problem.

**PUSHMEET:** So we look for problems which have a large impact.

**PUSHMEET:** We look for problems which require machine learning and AI.

**PUSHMEET:** And of course, we look at problems, in which there is feasible execution bias as I say, in the terms of we have data or some experience or a simulations to learn from. So , the role of machine learning AI and then feasibility in terms of the execution roadmap. So these are the things that we consider.

**PUSHMEET:** And of course, there is a lot of uncertainty associated with any of these directions, so we do a very comprehensive scoping exercise, which is like due diligence ,where we even take on sub projects to explore what this area is like, what has been done?

**PUSHMEET:** What are the baselines? Can we recreate the baselines? What is the true impact? We talk to subject matter experts in the case of, particular problems in the natural sciences. Sometimes even Nobel prize winners And ask them about their thoughts on the importance of problems.

**PUSHMEET:** What has been the key hurdles? So there's a lot of work that goes into the due diligence and scoping behind these decisions. And , this is not a decision that is taken top-down. The team collectively thinks about and evaluates all these different dimensions and then we rigorously discuss it, debate it, and then come to a conclusion as to how should we prioritize our researches?

**CRAIG:** How many people working in the group?

**PUSHMEET:** The core team in the science team is 50 people. But as you can imagine, at DeepMind, there is a large number of researchers, working on fundamental techniques in machine learning and AI. And a lot of those individuals are contributing to different research efforts. So we can significantly expand.

**PUSHMEET:** So if you look at all the people who are working on scientific projects at DeepMind, that number will be more than a hundred. But then if you look at the core team, which is currently just focused on doing the scoping exercise, framing it correctly, and so on that's less than a hundred.

**CRAIG:** What is the day in the life of a researcher in the science team, are they for months at a time focused on a single problem? Are they coding? Is it really conceptual and actually building models is left to a group of engineers? How does that work?

**PUSHMEET:** The life cycle of a project is certainly on the longer side at DeepMind because we take on such ambitious projects. Like protein folding, that was a four year project. And we're still continuing to work in that area. And theres similarly math

**PUSHMEET:** or quantum chemistry and so on, or even glassy dynamics. These are all very, long-term projects.

**PUSHMEET:** At DeepMind, we have this view that everyone does research, right? So even though we have roles like software engineers and research scientists, everyone does research in their own way.

**PUSHMEET:** Everyone is accountable for doing good research.

**PUSHMEET:** in addition to good research, The software engineers are accountable for thinking about, do we have the right infrastructure, the research engineers, in addition to doing good research, are accountable for whether we are taking all the best that has known about machine learning and AI towards that problem.

**PUSHMEET:** And, even the domain experts they are really thinking about, are we framing the problem correctly? Are we are we looking at the right goals? Are we looking at the right metrics in addition to actually coding and contributing to research. So everyone does research.

**PUSHMEET:** Everyone does model development.

**PUSHMEET:** Everyone does think about where we are going, but they have individual accountabilities in addition to that.

**CRAIG:** So you're doing research yourself, hands-on work. I imagine just managing 50 people is a job in itself.

**PUSHMEET:** If you're lucky enough to find people who are passionate about what the organization wants to achieve, then the actual act of management goes away. The need for management is reduced because these are individuals who are so passionate about science and the benefits that machine learning and AI can have in tackling some of the great challenges we are facing today, whether it's climate change or the pandemic.

**CRAIG:** One of the things that's struck me about your research from reading some of the papers and some of the blog posts is there's a real focus on the neural architecture to match the problem. How is that evolving?

**CRAIG:** On the one hand, deep learning research is striving toward the most general model possible. But on the other hand, you keep slicing into increasingly specialized architectures, specialized hardware. What direction do you see deep learning going?

**PUSHMEET:** Yeah, so that's a very good question.

**PUSHMEET:** And I think that this goes on to the approach we are taking.

**PUSHMEET:** The mission of DeepMind is basically on building techniques, which are general purpose.

**PUSHMEET:** They are applicable to many different areas. Even if you look at the architectures that we are employing for some of our very large efforts, like protein folding, what we are trying to do is bring the best experience for the learning system to leverage in terms of data, as well as all the information that has been gained by the scientific community in defining that problem, right? The underlying physics of the problem. And how do you integrate that? Now, some people think of machine learning as here's some data I'm going to apply certain techniques and out comes an answer.

**PUSHMEET:** Actually, machine learning should be thought of as an act of programming where you're telling the learning system about the problem at hand, and the information about the problem can be embedded implicitly in the architecture of the model, implicitly biasing the models towards learning the right thing.

**PUSHMEET:** Some of the inductive biases that our fundamental research at DeepMind has developed over the years, whether they are foundational neural networks and their extensions or transformers or auto regressive sort of models, they're all trying to build a collection of techniques or inductive biases that can be used to communicate with the learning algorithm as to what aspect of the problem is important. And in the future, one would hope that by looking at the problem and if provided enough data, the system would automatically learn what is the right inductive bias to employ?

**PUSHMEET:** At this point when you're really optimizing for one specific problem, we are trying to communicate to the system as to what are the different domains , specific things, which are important. And so the work on graph neural networks, if you look at the structure of transformer models that we have used in AlphaFold, it's trying to tell the learning system, what is the underlying physics. And it's not just relevant for say protein folding, but it's also relevant for anything that when you're reasoning about molecular systems, which are operating in 3D and have 3D interactions, which are not defined by neighborhood on a sequence, but they are defined by neighborhood in 3D, which you don't know at the get-go.

**CRAIG:** I do have one last question. In the problems that you're working on right now. Do you have a hierarchy in your mind, which are the most important for solving the problems in the world today?

**CRAIG:** Or are there any on the horizon that we should be paying attention to that DeepMind is going to tackle

**PUSHMEET:** there is of course, like sciences in the service of society. And if you look at particularly the challenges that society is facing today, of course there is the pandemic and understanding biology and what we can do and not just the pandemic, but also the health crisis that is happening across the board right now.

**PUSHMEET:** And it's not just pandemic but understanding biology. And how do you make sure that healthcare is available to everyone?

**PUSHMEET:** And then on the other hand, climate change and the things associated with it. What can science do for clean energy? So problems like fusion, problems like new materials for better superconductors, for a variety of applications. So we are driven by scientific problems that would impact society and would help society address some of the challenges that we are facing today. And I particularly see biology and some of the problems that come out of the climate crisis as the one's to invest a lot of effort in.

**CRAIG:** That’s it for this week’s episode. Again, I want to thank [ClearML](https://t.clear.ml/eye_on_ai) for their support. Please take a moment and visit [clear.ml](https://t.clear.ml/eye_on_ai) to see what they offer. I think you’ll want to give them a spin.

I want to thank Pushmeet for his time. We’ll be watching his team’s work.

In the meantime, remember, AI is about to change your world, so pay attention.