**Sergey:** 0:00

Imagine someday building a home robot that can perform a variety of tasks in your kitchen. Then it becomes a lot more than just a perception problem. Then you really need to be able to learn the individual manipulation skills you need to be able to generalise very broadly.

**Craig:** 0:12

Hi, I'm Craig Smith and this is Eye on AI. Today I talked to Sergey Levine, an associate professor at the University of California, Berkeley, who does research at the robotic artificial intelligence and learning lab at the university and is pushing the boundaries of what AI control of robots can do. Sergey talked about some of his recent work in reinforcement learning and the aggregation of data sets from robots around the world to help train a model that can generalise across different kinds of robots. It's exciting research into embodied AI, bringing the transformative technology out of the computer and into the real world. I hope you find the conversation as fascinating as I did. So, Sergey, could you start by introducing yourself?

**Sergey:** 1:11

So I'm an associate professor at UC Berkeley, I previously did my PhD at Stanford, and I also spend one day a week at robotics at Google, where I also work on robotic learning. My research concerns, of course, robotics, but also many other related techniques in machine learning, reinforcement learning. Lately, my group has also been doing work on various things related to reinforced learning for language models, computational design, things like this, and other aspects of decision making.

**Craig:** 1:43

Everyone's talking about world models and they're combining world models and language models. Are you working on world models at all? What's your view of that?

**Sergey:** 1:54

Yeah, I guess there are a few things I could say about it. So generally, if we want to control robotic systems, there are a number of ways that machine learning can enable that. One very simple way is imitation learning. Imitation learning amounts to taking demonstrations, typically provided by a person controlling the system, and then imitating those demonstrations to try to produce an agent. Robots can work for lots of other things. Arguably, language models are just giant imitation learning machines because they're imitating how humans generate text. There are lots of other ways to do it.

**Sergey:** 2:32

So a world model is essentially a dynamics model that represents how the environment will evolve in response to the agent's actions, and we can learn that from data too. Typically in reinforced learning that's referred to as model-based RL. So model-based RL means you train a model of how your environment behaves and then you use that model to figure out how to act in the world. And it's a very old discipline actually. In fact, the first learning-based control methods, before model-free RL became such a popular thing, were actually model-based RL methods. Some of the earliest neural network control methods actually used dynamics modelling. And again, there are a million different ways to instantiate this. You could instantiate dynamics models or world models by, for example, taking image observations and doing video prediction. You could also instantiate them by learning non-reconstructive representations, or representations that, roughly speaking, capture the state of the system without necessarily grounding it back into pixels, and then predict that. So there's a lot of different approaches for doing this.

**Craig:** 3:35

Recently I've been talking to Wave about their Gaia model and I've seen the videos. But they have that model built into a controller, connected to a controller, to operate an autonomous vehicle. What's different about that structure or architecture from what you're working on with reinforcement learning?

**Sergey:** 4:01

I don't think there's that much I could say about it, because I don't know how their system works. I mean, I've seen the public material, same as everyone else, but I don't really have any insight about the details there. Maybe one thing I would say, though, that most methods for learning-based control don't necessarily need to predict the raw pixels that the robot's camera will observe in the future. That is one way to do it, and there's a lot that can be done with that, but I think that the more significant distinction is actually how well we're able to utilise data to then produce more optimal decisions, and going through prediction is one way to do it, and you could predict pixels, which is what video prediction models do. You could also predict outcomes or rewards, which is what value functions do. At the end of the day, they're actually not all that different, and maybe a bigger distinction as to whether you can get a system that really works in the real world is what data is trained on. For example, if you want robotic manipulation systems that will actually work in broad, open-world environments, you need to train them on data for broad, open-world environments. So a lot of what I'm actually concerned with in my research is how do we develop learning-based control techniques that can use large amounts of data, and how do we figure out what data sets we can acquire to get really generalizable? In my case, often robotic manipulation skills, but also robotic navigation skills and things like that Systems for manipulation for things like warehousing.

**Sergey:** 5:34

Oftentimes those problems can be to a large extent reduced to perception problems. So if you structure your environment in the right way, then as long as you're able to detect where the objects are, then you can use hand design strategies for tackling that. That tends not to work so well. If you want to take the robotic system into more open-world environments Like if you imagine someday building a home robot that can perform a variety of tasks in your kitchen, then it becomes a lot more than just a perception problem. Then you really need to be able to learn the individual manipulation skills and you need to be able to generalise very broadly.

**Sergey:** 6:05

So maybe one thing that I could discuss here that could be relevant is a project that we actually did recently. This was actually a collaboration between Google, Berkeley and several other universities on trying to see how we could get robotic controllers that actually generalise across different robot morphologies. That's actually really important because if a lot of this comes down to data, then, it's very difficult to get the kind of breadth and diversity of data from a single robot that can allow the kind of broad generalisation that you want from a home robot. But if you can pool data from lots of different robots, then maybe you can actually get that kind of coverage. And, furthermore, if you can actually do that and you get a system that generalises across robots, then you get something really cool which is, in principle, someone could put together some new robotic system and then plug this kind of robot brain into it and immediately get something that can control that robot. Now, the work that we did so far on this wasn't really that concerned with building better models so much as just getting this diverse data set and just applying the kind of standard techniques that we had already developed in previous works and that actually worked decently well. So this project was called RTX and the idea there was that we got data from In the end it was 34 different research labs.

**Sergey:** 7:19

Google was one of them, Berkeley was. Actually there were two labs in Berkeley that contributed to this and then we trained a model on this to perform basically language-conditioned manipulation tasks.

**Sergey:** 7:29

So I think you give the robot an instruction, pick up the tomato and put it in the bowl, and the robot is supposed to do that. And then we took this model and we handed it off to the different labs that contributed data and had them tested in comparison to whatever model they had for their research, basically trained on their own system, and the multi-robot model actually was, on average, about 50% better in terms of success rate, and that's actually pretty interesting because this was competing with whatever each lab's individual system was and presumably third-good researchers. They built something that works pretty well in their system. Now, this was actually an imitation learning approach, language-conditioned imitation. I think that, whether it's imitation or prediction or world modelling, I mean, I think many of those techniques could be made to work. I think the higher-order bit here that I wanted to get across is that by actually getting this data set, you could actually get a system that you could plug into all these different robots and actually get good results out of it.

**Craig:** 8:16

Hmm, that's fascinating. That model was being trained on data sets from all the various participating labs.

**Sergey:** 8:25

Yeah, so in these experiments we were not testing whether it could generalise to a new robot. That's a very exciting frontier for this, but that's still in the future. This was just trying to answer the question: if you include data from the other labs, does that one lab's robot get better? Now, of course, if you're sort of in the minority, if you're one of the groups that contributed a relatively small amount of data, you would expect to see comparatively more benefit from everyone else. Interestingly, actually, even the majority contributors saw a lot of benefits.

**Sergey:** 8:52

So probably the largest data set at about 100,000 trials was from Google's own robot, the mobile base that we used in a lot of the robotics research there. With that system, we were able to actually test it on various tests. We have this kind of test suite of difficult queries. They're actually meant to be queries that require synthesising pretrained knowledge from the web as well as good instruction following ability, so these require spatial reasoning, things like that, and on the hardest test, we actually saw three X improvement in performance over just using the Google data set. Now that's actually, in my mind, pretty profound, because the Google data set was very carefully curated, collected by basically professionals that were collecting robot data, and the fact that including all these additional sources of data from a long list of academic labs actually led to that much improvement really suggests that there is something kind of magical that happens when you combine enough data from enough different sources. Yeah, so for these experiments we were actually passing the model around. Okay, the data set is public now.

**Sergey:** 9:54

So anybody could take that data set and download it and train their own models. And we actually have an ongoing effort at UC Berkeley that my students are For that initial experiment. It was just the model weights. Hmm, that's fascinating. Just the model weights.

**Craig:** 10:09

So the architecture of that model was being replicated in each lab. They weren't using their own model.

**Sergey:** 10:16

Correct, yeah, so it was exactly the same model with exactly the same weights, that had to drive all of the robots in all the locations. Yeah, and that is actually, if you think about it, it's a very non-trivial thing, right? Because all the model gets to see is what the robot receives through the camera and has to figure out that. Okay, now I'm driving a U-shaped robot, A UR-10 industrial robot, versus now I'm driving a low-cost WTOX robot. Or now I'm driving a Franca or the Google robot and adjust the controls accordingly.

**Craig:** 10:44

When I was at the lab, I recall that you had robots that were network, so the learnings from one were updating a central brain and that then was controlling the other robots of each robot. Have you done broader experiments like that, very much like this?

**Sergey:** 11:05

Yeah, yeah, I'm glad you asked that. So this was a lot of what we were trying to do for over the last five years actually, and in some ways, this multi-robot training effort it's kind of partly it came about kind of an acknowledgement of the limitations of this kind of arm farm approach. So getting lots of robots in a room is great if you want to prototype, let's say, reinforcement learning algorithms, but if you really want broad generalisation out of it, they can't all be in the same room. So you really need to get much better coverage of the world, and by aggregating data from robots in many different sites now, you can get much better coverage. Now this is still a prototype for what might be a larger system, because these are still data sets collected by researchers essentially doing science experiments. So you could imagine that in the future, the aggregation wouldn't be across different research labs, it would be across different deployed robots.

**Sergey:** 12:01

Now that, of course, is a much more complex undertaking that requires more than just science. It also requires some kind of organisational effort, consensus from companies and so on. But that, I think, is kind of the real thing once that comes about and you could imagine a future where the data streams from a variety of different deployed robots in lots of locations are agglomerated and then used to train one centralised robotic brain, which can then be handed back to these robots to improve their performance. And the key thing that we want to take a risk with this project is just if you do this at any scale, you know, even at the scale of academic labs can you even get a policy that can drive all the different robots? Because if that's not possible, then aggregating heterogeneous data wouldn't work and we would need to somehow figure out standardisation. Standardisation is hard, so what we know now is that we don't have to worry as much about standardisation.

**Craig:** 12:50

Yeah, because, yeah, this model, then the weights are being passed around and they're controlling different form function robots, right, I mean? Or were they all just variations?

**Sergey:** 13:00

So in these experiments the robots were all arms with parallel jaw grippers. We are right now experimenting with generalisation across single arm and bimanual systems. At some point in the future we'll also look at multi-fingered systems, things like that. So far the truth in advertising it's one arm with a parallel draw gripper. They're just different brands of arms. Now they do vary a lot. So the small scale hobbyist widow X arm is maybe 50 centimetres long, kind of smaller, with a weak gripper. The big, your 10 robot, is an industrial robot meant for manufacturing, quite a bit bigger, beefier, has a stronger motor, stronger gripper, that sort of thing. So there's variability. They're still the same sort.

**Craig:** 13:43

Right, and the model that you're training on this aggregated data is the reinforcement. Can you describe that model?

**Sergey:** 13:50

We actually trained two models. One was based on the RT1 model which was developed at Google last year. The RT1 model is basically a transformer that reads in language instruction, commands, an image, and then it outputs discretized, tokenized actions. So it's kind of almost like the most obvious way to design a transformer based policy. The second model was the RT2 model, which is a more recent development, which actually uses a backbone from a pre-trained vision language model.

**Sergey:** 14:23

So vision language models are trained to look at images and output responses to textual questions. So you give it a picture and you say like is there a dog in the picture? And it will produce some text to answer that. And then we took this vision language pre-trained backbone and then further fine tuned on robot data to output robot actions in response to robot observations. So you can sort of think of it like the VLM has a number of tasks that it can do. It can answer questions, it can produce captions. Now there's one more task which is given a robot instruction output the actions for the robot. Now that's a much more powerful model because it has all that internet knowledge baked into it from the vision language model pre-training and that's the one that we used for the more complex queries with the spatial relations and things like that.

**Craig:** 15:03

Is most of your work on the data side or on the model side.

**Sergey:** 15:08

Well, it's really both and to some extent they also go hand in hand because, depending on what your algorithms are able to handle, that will inform the kind of data that you need to get. For example, a lot of the more algorithmic work that my lab does these days is concerned with techniques for offline reinforcement learning.

**Sergey:** 15:26

Offline reinforcement learning is basically a way to take data and produce more optimal policies. So imitation learning methods. They take in data and they produce policies that reproduce the behaviour in the data. Offline RL methods take in data and attempt to produce behaviours that are better than the average behaviour in the data. So intuitively you can think of it as using data to figure out what options are available to you and then selecting the best among those options. In fact, methods that use world models, as we discussed before, can be seen as offline RL methods because typically the way they work is they train the world model on existing data and then they use it to extract better control strategies than the typical thing that was demonstrated in the data set. But there are also other ways to build offline RL techniques that don't rely on world models, that rely on value functions and things like that.

**Craig:** 16:06

Where do you think the research is heading, because everything's moving so fast? Is there for robotic control? Do you think that the research will settle on one architecture and then there will be different flavours of that architecture, but everyone will agree that this is the best way, and then it's a matter of training, generalising across robots and networking data? Or do you think that there will be a series of models that will be used for various functions?

**Sergey:** 16:43

Yeah, good question, so I'll give you an answer. It's a slightly aspirational answer, so maybe this is more like where I wish things were headed. I don't know if this is necessarily where things will head, but I think it's very important for robotics to kind of adopt a paradigm where we are in the habit of having reusable models, where, just like in computer vision and NLP, if a researcher produces a good model, other robotics researchers should be able to use it. Now, that might seem like a very obvious thing, but this is not actually how robotics works today. Most robotic learning research, the artefact that is produced is not actually the model, it's the code or the paper or the insights. The models themselves are almost never portable, never mind across labs, even across different locations in the same lab, different times of day in the same lab, that sort of thing.

**Sergey:** 17:35

And I think we really need to move that towards a setting where we have models that are trained on datasets that enable generalisation across different locations and systems, different objects, that sort of thing that we can then give to other researchers, other practitioners that will also run on their systems, and once we've got a good flow for doing that, maybe using things like this RTX dataset that has multiple robots, maybe using some other data, but something where we can just get on the habit of doing that.

**Sergey:** 18:03

Then we can actually make more progress as a community towards shared, generalizable systems. Now, until that happens, there's absolutely no question about whether people will use the same architecture, the same model, like if they can't even share anything across, then that won't work. But once we can share something and probably the key to that is a dataset that enables that then the community can figure it out, like, maybe at that point perhaps it'll be realistic to have a single pre-trained backbone, like the Lama model in natural language, processing an analog to that and robotics, and then people can build on top of that. Or maybe there will be several such things. Maybe there will be a few big ones, that kind of the big, well-equipped labs produce that others will then build on. But before we get to any of that we need to just get in the habit of actually building models that others can run.

**Craig:** 18:50

The other side to robotics is just the hardware, and I was talking to a guy the other day who was talking about where robotic control systems are heading, and he was. He's not a roboticist or an AI researcher, but he was waxing very optimistic about, you know, there being household robots within three to five years, and that sounded unlikely to me, because just the hardware alone is not at least the hardware that I've seen is nowhere near being able to do, you know, releasing it into an unstructured environment full of randomness. Do you think that the hardware is moving along with the AI or is it lagging?

**Sergey:** 19:44

That's a good question. I think that a very important part of that question is just what kind of hardware we need. I think to a large extent, learning methods should actually lower the bar for the hardware that's necessary. Basically, the exercise you can do is you can get one of these like little trash-bicker devices and see what kind of tasks you can do around the house with it. I mean, obviously it's very limiting, so there's some things you can do you couldn't do, but there's also a lot of things you could do with it. Certainly you can, you know, tidy up the floor, put things in different locations in the kitchen. It's actually kind of surprising how much a relatively primitive robotic system can accomplish.

**Sergey:** 20:24

So there's a very nice work out of Professor Chelsea Finn's group that I also helped with a little bit, by a student named Tony Zhao who developed a bimanual robotic system out of two low-cost robots from Trostin Robotics. So these are not even the fancy industrial arms. These are basically very fancy hobbyist robots. So they cost about, I think, $5,000 each, and most of his kind of cleverness in his research was in devising a very convenient tele-operation system, a tele-operation rig that he could hold with his hands and control this fairly cheap bimanual system and he would demonstrate all sorts of very complex behaviours. You get this thing like putting a shoe on a foot, using tape to tape down a box, things like that, and then you know the learning methodology that could produce the autonomous policy was well designed but not particularly profound.

**Sergey:** 21:24

It sort of used state-of-the-art transform-based techniques but didn't really have any particularly surprising innovation. The key to it was really building a really good tele-operation rig that allowed him to produce those behaviours and then a very high-quality engineer to then get that up down to policy. So this is called the Aloha system, you know, for those who are listening, I encourage you to check it out and it probably gives some idea of what even very primitive hardware is capable of if it's equipped with the right data, the right kind of tele-operation rig to provide that data and kind of good bread-and-butter modern machine learning techniques. Now that's still not going to do everything around the house, but I suspect that for folks that watch these Aloha videos it'll kind of maybe slightly change their mind in terms of the kind of hardware we really need for everyday tasks. So probably there is still some innovation, but it might actually be less than you think.

**Craig:** 22:15

That's interesting. And then the controller side, the AI side, the model side, is it? I mean, if that is adequate, that hardware, how much more improvement is needed on the control?

**Sergey:** 22:30

side. That's a complex question because that's probably very heavily dependent on the required bar for robustness and the degree of generalisation. So in some ways that sort of parallels the autonomous driving story right, like if you wanted to build an autonomous car that could, you know, succeed in like 90% of cases, well, that's probably something that we've had for over a decade. But if you want an autonomous car that will succeed, that will avoid catastrophic failures, with enough robustness that you could just deploy it on any road in any city, just dealing with all those tail cases, that's still an open problem and I think with home robots it's going to be the same way that if you want to lop off the bulk of the things and the bulk of the situations, maybe that's not quite there yet, but I think that it's reasonable to imagine that we get there soon. But how long it takes to get that long tail fully figured out, that's a much more complex question.

**Sergey:** 23:22

I think that one thing that's pretty interesting is the degree to which vision language models have progressed over the last really over just like the last 12 months and that's especially relevant for robotics, because, while the way that vision language models are typically used is more for you know, kind of perception, traditional perception tasks, question answering that sort of thing, the ability to reason about visual observations, perform inferences about spatial arrangements of objects, that sort of thing that is something that is likely to translate into better robotic capability, and, because generalisation is one of those big challenges I mentioned the slump tail issue I think there is a lot of reason to be optimistic about the potential for those models to eventually improve the robustness of robotic controllers as well.

**Sergey:** 24:11

People are talking about combining language and vision, or I should say language and world models, into agents that can reason, plan and take action. That sounded to me very much like robotic control.

**Sergey:** 24:27

I guess what I'm asking is that research and the people who are in robotic control, research on different tracks, the answer is a little bit complicated, but maybe the short version is that, yes, it's closely related to a lot of robotics problems. In fact, there's plenty of work in robotics on using language models for essentially constructing plans and then connecting those plans up to some kind of control mechanism that can bring them about. Now, probably, this stuff started maybe roughly two years ago. Probably one of the more well-known works in this area is the Seikan paper from Google, which used a language model to plan long horizon behaviours for robots. Initially in this field, one of the big challenges that people were concerned about is how to connect up the language model to perception and action, because standard language models have to operate on symbolic representations of the world, so you have to take those symbolic representations and somehow weld them on to rich sensory perception and complex actuation. Now, initially the way that this was done was kind of along the lines of what you described, by trying to construct some sort of joint planning procedure that would figure out both a probable sequence of symbolic steps, essentially language and the corresponding behaviours that would bring that about. There's actually a paper from one of my colleagues from Skult, grounded Decoding, which proposes a Bayesian filtering approach to doing exactly that. That said, something that we've seen over the last maybe like six to nine months is that increasingly, with vision language models becoming more powerful, a very appealing alternative instead of doing this is to actually train models end to end to solve the entire problem. Now those models can still be doing planning.

**Sergey:** 26:21

If you have a vision language model that outputs text and also outputs actions, you can do essentially the analog of chain of thought prompting. You can say okay, here's some complex problem and produce steps for solving that problem, and then, once you've produced those steps, then produce the actions and that works. So you could tell a robot okay, like, make breakfast and also get to make breakfast, I need to do this and this and this, and then, for the first step of that, it'll try to output the actions. So that's a viable way to use vision language models, but then it's still. You would still end up with one model that does that, and that's very desirable, because if you have one model, then you don't need to deal with this problem of trying to somehow stuff visual observations into symbolic representation to then pass into the language model. Basically, instead of designing that interface by hand, it emerges naturally through joint training of the whole thing.

**Sergey:** 27:09

So this is actually the principle in which the R2-2 model works, and one of the examples there that illustrates this kind of chain of thought style approach is as we asked it. We wanted to intentionally construct a scene where the correct behaviour is a little bit non-obvious. So we had a scene that had some common household items and had some tools with the wrong kind of tools so it's supposed to hammer in a nail. There was no hammer, but there's a rock and we asked, "Okay, you need to hammer in the nail, what should you do? And then it figures out that you should pick up the rock. It actually says rocks and then it goes and exits to the corresponding actions. So now that's very primitive planning, right? So it's kind of more semantic inference than planning. But these things are in their infancy. I think they'll progress a lot more over the next few years.

**Craig:** 27:55

Do you? In the last five years, which is about the time I think since I spoke to you, has the progress in your field specifically mirrored the progress in generative AI?

**Sergey:** 28:14

I think that progress in robotics always to dust into lag behind everything else, because when we figure out effective learning techniques, then we then it's always a longer journey to go from kind of conceptual method to product, to a small scale prototype, to larger scale prototype, because with generative models, well, you can harvest lots and lots of data off the web, so the lag between developing a method and then scaling it up to internet scale data is typically relatively short.

**Sergey:** 28:46

With robots that's usually not the case. So while certainly modern advances in generative modelling have made a big impact on robotics and there's particular very interesting adaptations of those techniques that combine with reinforcement learning, planning and so on, I would say that so far we have a lot of good indications of the potential for these things, but we don't have the kind of large scale prototypes that have been produced, for example, for diffusion models, for image generation or for language models. And I think the key there is actually getting these kinds of reusable models with large and diverse data, so that would make it possible for us to produce these larger prototypes.

**Craig:** 29:23

Yeah, so what's next in your lab?

**Sergey:** 29:27

Yeah, so one of the things we would like to do is provide the community with pre-trained models, now that we actually have a data set to work with that can be easily adapted to a variety of downstream applications. So not just a model that can do anything and that's maybe too ambitious of a goal but at least a model that can be adapted to do anything. So if you could imagine, let's say, a model that is pre-trained to take in language, take in maybe goal observations, other forms of commands, and produce outputs for a variety of different robot embodiments not with the goal necessarily of solving every problem, but providing a really good initialization. So that's somebody that has a particular specific robotic system with a particular desired formulation of their task, a particular objective. They could take this and, with a much more modest amount of data, adapt to their problem. And I think that now that we actually have good multi-robot data sets and fairly mature techniques in terms of how to train models with variable inputs and outputs, we're actually just about ready to do that. So our first prototype for this should be coming out very soon. But this is going to be the first step.

**Sergey:** 30:30

From there, a lot of what we have to investigate is what does the life cycle of such a system actually look like? What are the right techniques for efficiently fine-tuning robotic foundation models to particular domains, to different morphologies, different commands and so on? And there's probably actually a lot of interesting questions to be answered there. For example, robots can collect data autonomously, so could you, for example, have an autonomous fine-tuning procedure based off of one of these pre-trained models? Could you have a fine-tuning procedure that respects safety constraints, things like that? So there's a lot of interesting questions that we could answer once we have that base model all set up.

**Craig:** 31:05

There's a lot of people I've been talking to about the proprietary, open-source debate. With regard to generative AI In robotics, is there an analogous situation where there are enterprises that have tremendous resources? I mean, the robotics are not as compute-intensive as the models you're talking about. Is that right, and so is it more equal at what's happening in industry and what's happening in research?

**Sergey:** 31:38

Yeah, it's complicated. So certainly compute constraints are an issue right, especially once we go into vision language models. The most effective vision language models are actually the largest models out there. So the largest version of the R2-2 model, for example, is 500 billion per amperes, so very much in the same ballpark as the largest models out there. Of course, you can do a lot of experiments at a much smaller scale, and that does make it somewhat more accessible.

**Sergey:** 32:03

In terms of data. It's kind of interesting. There are definitely companies that have large numbers of robots deployed. Those are not necessarily the companies that have the most interesting data, though, because if they're deployed in a warehouse, it's mostly grasping of items and maybe in some ways, the open data from researchers is actually more interesting. The picture changes somewhat if you go into mobility, things like autonomous driving, like yeah, there's going to be big industrial players with their own proprietary stuff, but even there, like data sets constructed from dashboard mounted cameras that are out there now are actually very large Certainly not as large as what Tesla or Waymo has, but substantial. So I think you're right that some of the proprietary advantages are not as large, but it's kind of. Maybe the more pessimistic take on it is that it's because no one has the data. So the companies don't have the data because no one does.

**Craig:** 32:58

the control of an autonomous vehicle and the control of a robot arm or some other form factor, but are they different fields? I mean, when you're working on these models, are you also thinking about their application in autonomous vehicles?

**Sergey:** 33:17

So traditionally these are extremely different problems, but what we're increasingly seeing is a degree of consolidation in the sense that very similar building blocks can be reduced. So I think actual autonomous driving is maybe one of the tougher things because of all the constraints and regulations and all that stuff. But for small scale mobile robots like Think, like Drones, sidewalk Robots, et cetera, we already have research projects where we've developed vision based navigation policies. For these things they use essentially the same exact architecture as what we use for the robotic manipulation problems, and a very natural next step is to actually combine, not just have the same architecture, but literally the same model.

**Sergey:** 33:55

In principle, at this point, there isn't really any technological obstacle to doing that. Now, of course, there's a lot more to driving, let's say, an autonomous car than just avoiding obstacles and reaching the destination. You have to put in a lot more knowledge, constraints, all this other stuff, and that probably is rather specialised. But my hypothesis is that we'll probably actually see a lot of consolidation around the same basic building blocks for the core perception action system within these things, and then maybe where they would differ is the kind of planning layer that sits on top of that and then directs it in terms of what to actually do in a given situation.

**Sergey:** 34:29

In your work? There is a pull for academics because of the computer constraints and money and salary and that sort of thing to pull people into the industry. Are you working straddling academia and industry? Are you firm?

**Sergey:** 34:51

Yeah, so I spend 20% of my time working with Google DeepMind. I think that, in terms of the degree to which an industry researcher or academic researcher in robotics is more or less appealing or progressive, I think probably it's a little more, I would say, tilted towards academia than, for example, natural language processing or vision, and maybe in part it's because there are a lot more of the big questions to figure out before things actually have revenue, so to speak. You could build a language model or vision system that provides an actual business case today, whereas analog and robotics are probably still a few years out. That said, I do think that there's a lot of rapid progress and certainly a lot of students from my group are excited about starting companies based on technologies they're developing and things like that. So I think we'll see that catching up in the near future

**Craig:** 35:47

And you think that this is the year that AI hit the public sphere and people confuse robotics with AI all the time. Will the day come I mean, obviously the day will come, but when do you think the day will come that there will be some commercial application or open source application that is adopted by the public, that people will suddenly be talking about robots as opposed to AI?

**Sergey:** 36:23

Yeah, that's a complex question because I think that if I had to guess, I would guess that a lot of what's needed beyond the core technologies is a pretty substantial upfront investment to overcome the activation energy to get somebody about to be practical. And that's not too unprecedented in the sense that more or less the same thing happened with language models. The core technology for next token prediction is pretty old. What was needed to develop the kinds of technologies and products that really capture the public imagination is a large investment of effort into engineering them really really well and curating, collecting and assembling the right data sets to get them to work really well for a thing that basically anybody could grab and use. That's partly.

**Sergey:** 37:14

There's a scientific question there, but a lot of it is really sort of organisational economics kind of questions, and the trouble with those things is that they're hard to predict because they have more to do with the point in time at which people decide that it's time to lay out those big resources to make it happen rather than just predicting when the technology will evolve. So the technology might evolve steadily, but then the inflection is really the resource allocation, so I can't predict when that would happen. If I had to bet. My bet would be closer to five years than to 10, but I'm not sure.

**Craig:** 37:49

Has been. This threat debate has created a lot of acrimony in the community. Do you have a view on that, or is your territory removed enough that you don't engage in that?

**Sergey:** 38:08

Yeah, it's a complicated question. I tend not to prefer not to get too much into discussions like that Because I'm not really sure exactly how things will go, and I think that, partly perhaps as a roboticist, I Maybe tend to be a little more pessimistic about where we're at in terms of overall AI systems. It's hard to imagine that an AI system that can't control a robot to do basic things that are easy for humans would be all that capable overall, but the stuff is hard to predict. I think that maybe the one constant in AI research is that people are continually surprised by the things that turn out to be easier than imagined and also the things that turn out to be a lot harder than imagined. If you go back a couple of decades, it'll be pretty shocking to think that, for example, artists and writers would feel threatened by AI systems long before, like the gardeners and cleaners would, but that is the world that we live in today. Maybe that tells us to be a little humble about our predictions.

**Craig:** 39:14

Yeah, that's right. The governments around the world are very focused on regulation of generative AI specifically. Is there regulation or governments looking at robotics, or AI and robotics? In the same way, is there government support? There's been a lot of talk about providing compute resources to research and smaller companies so that that's not held within these big tech companies. Is there that kind of talk in robotics that the government should or could provide more resources to accelerate research?

**Sergey:** 40:04

Yeah, there's certainly plenty of talk about that. It's typically not, from what I've seen, not something that tends to separate out robotics or AI versus other things. There's certainly talk about it. I haven't seen a great deal of action yet, but I imagine it's something that moves slowly. I don't think I would say anything differently than any other AI researcher in that respect. I don't think that, from what I've seen so far, there's anything that treats robotics in a particularly special way in that regard. But yes, this is a big issue and it's probably something that we, certainly in the United States, need to think carefully about how we're going to maintain our technological edge and how we're going to allocate the resources that are necessary for that.

**Craig:** 40:46

And that leads me to another question, because I spent a lot of my life in China. Where is China in this research? Do you think they're ahead, behind?

**Sergey:** 40:59

I'm not sure exactly. One thing I will say is that I think that researchers from China, from Chinese universities, have been very successful across all areas of AI, including robotics, and certainly a lot of really interesting research increases in IDUCI coming out of China. For example, when we were doing a lot of our data set collection work, we were actually very surprised to learn in the middle of it that there was a really amazing data set that was released by some researchers from Shanghai that was comparable in size and scope and diversity to the one that we were collecting, and that was wonderful. They released it open source. I talked to them on the phone. They had really interesting thoughts about what they wanted to use it for, so I am seeing a lot of uptick in terms of the quality and the kind of results that are coming out.

**Sergey:** 41:47

The other interesting thing is that actually a surprising number of advances in hardware have actually been enabled by companies out of China. For example, one of the most widely used platforms for quadrupedal locomotion research is produced by a company called Unitary from China, and I think a lot of the things that make that platform so appealing are that it's relatively simple, it's affordable and it's designed in a way that makes it easy for researchers to get into the guts of it, and that also, in my opinion, has actually been a very good thing, because, while we might be concerned about competition and all that, in the end it's actually accelerating research here in the United States. So that's what I've seen so far, I mean without trying to make any value judgments on what's good or bad. It seems like there's a lot going on.

**Craig:** 42:32

That's it for this episode. I want to thank Sergey for his time. If you want to read a transcript of today's conversation, you can find one on our website, EYEONAI, that's E-Y-E hyphen on AI. In the meantime, remember the singularity may not be near, but AI is changing your world, so pay attention.