EPISODE 75

[INTRODUCTION]

[0:00:10.4] SC: Hello and welcome to another episode of TWiML Talk, the podcast where I can view interesting people doing interesting things in machine learning and artificial intelligence. I'm your host, Sam Charrington.

A bit about the show you're about to hear. This show as part of the series that I'm really excited about in part because I've been working to bring it to you for quite a while now. The focus of this series is a sampling of the really interesting work being done over OpenAI; the independent AI research lab founded by you Elon Musk, Sam Altman and others.

A few quick announcements before we dive into the show. In a few weeks, we'll will be holding our last TWiML online meet up of the year. On Wednesday, December 13th, please join us and bring your thoughts on the top machine learning and AI stories of 2017 for our discussion segment.

For our main presentation, former TWiML Talk guest, Bruno Goncalves, will be discussing the paper; Understanding Deep Learning Requires Rethinking Generalization by Chiyan Zhung from MIT and Google Brain and others. You can find more details and register at twimlai.com/meetup.

Also, we need to build out our 2018 presentation schedule for the meet up. So if you'd like to present your own work or your favorite third-party paper, please reach out to us via email at team@twimlai.om or ping us on social media and let us know.

If you received my newsletter, you already know this, but TWiML is growing and we're looking for an energetic and passionate community manager to help manage and grow programs like the podcast and meet up and some other exciting things we've got in store for 2018. This is a full-time role that can be done remotely. If you're interested in learning more, reach out to me for additional details. I should mention that if you don't already get my newsletter, you are really missing out and you should visit twimlai.com/newsletter to sign up.

In this episode, I'm joined by Dario Amodei, team lead for safety research at OpenAI. While in San Francisco a few months ago, I spent some time at the OpenAI office during which I sat down with Dario chat about the work happening at OpenAI around AI safety. In our conversation, Dario and I dive into the two areas of AI safety that he and his team are focused on; robustness and alignment.

We also touch on his research with the Google Deep Mine team, the OpenAI Universe tool and how human interactions can be incorporated into reinforcement learning models. This was a great conversation, and along with the others in this series, this is nerd-alert worthy show.

A quick note before we jump in, support for this OpenAI series is brought to you by our friends at NVIDIA, a company which is also a supporter of OpenAI itself. If you're listening to this podcast, you already know about NVIDIA and all the great things they're doing to support advancements in AI research and practice.

What you may not know is that the company has a significant presence at the NIPS conference going on next week in Long Beach, California including four accepted papers. To learn more about the NVIDIA presence at NIPS, head on over to twimlai.com/nvidia and be sure to visit them at the conference. Of course, I'll be at NIPS as well and I'd love to meet you if you will be there, so please reach out if you will.

Now, on to the show.

[INTERVIEW]

[0:03:46.4] SC: All right, everyone. I am here at the OpenAI Offices and I am with Dario Amodei. Dario is a team lead for the safety research team here at OpenAI. Dario, welcome to the show.

[0:03:59.0] DA: Thanks for inviting me.

[0:03:59.9] SC: Absolutely. It's great to have you on. Why don't we get started by having you tell us a little bit about your background and how you got interested in AI, in AI safety in particular?

[0:04:09.9] DA: Yeah. Actually, my background was in computational neuroscience. I did a Ph.D. in biophysics and computational neuroscience. Have always been interested in AI and how intelligence works, but I felt like a lot of our AI system 10 years ago weren't working that well, and so I decided I wanted to study the brain. Then of course the deep learning revolution came around and I looked at it and I said, "Oh! These system are actually starting to work. I want to be part of this. This now seems like the most interesting thing." I ended up working in Andrew Ng's group at Baidu for about a year and then I worked at Google for about a year, did kind of variety of stuff; speech recognition, natural language processing, and then I came here.

The way I got into AI safety was one of the things I kind of noticed about in neural nets is there's this mixture of they're very powerful, but they can be very opaque and unreliable. When I was developing speech recognition systems, you train a system with an American accent and then if it hasn't been trained in a British accented speech, you give it some British accented speech and it gets kind of totally confused.

This mixture of power, opacity and kind of very unpredictable failures and weakness is kind of what made me think that we need to be careful to make sure that these technologies do what we want them to do and don't do something unpredictable or even dangerous.

[0:05:32.9] SC: Nice. Maybe we can get started by having you give us a little overview of kind of the research that you've been working on, like the broad brushstroke overview of the research that you've been taking on.

[0:05:46.6] DA: Yeah, sure. I would say there're kind of two general areas that I think about when I think about safety. One of them is what you might call robustness, which is the problem that you have when you train a machine learning system like a neural network on some problem like speech recognition of self-driving cars. Then you put it work in some actual context and the distribution of inputs that you're facing changes in some way. For instance, the change in the accent, the change in the speech accents, the change in conditions of a self-driving car or if you

have a reinforcement learning agent exploring the environment when it sees changes, so dealing with that.

We've done a little bit on that. We've done kind of more on sort of the second thing, which is what we call kind of like alignment with human goals. I think a bit concern particularly as AI systems get more and more powerful is making sure that the AI systems have a sophisticated high-level understanding of what humans want them to do and that we can already see examples today where AI systems kind of it's much easier to have a simple goal than it is to have a complex goal, and I can get into a little more detail of why that's true for today's AI systems. That kind of really sets the stage for systems to kind of go off and kind of maniacally pursue very kind of pathological simple goals. I think we should instead have AI systems that are as good at understanding what humans want them to do as they are accomplishing whatever task they've decided it's important to do.

[0:07:23.1] SC: I think one of the best examples of that was from — I think it was some of the initial work that you did with Google on this where the example was you've got a robotic housekeeper and tell her to clean the room. It's like you don't want the robots sweeping all the dust under the rug.

[0:07:39.7] DA: Yup. Yeah, it's kind of an example of what we call, or just a term from economics actually, Goodhart's law, which is that once a metric becomes a target, it seizes to be a good metric. What that means is if you have a way to kind passively measuring something, it seems good, but then if you optimize it really hard, you might get something that you don't expect.

[0:08:01.0] SC: Isn't that all we're doing in AI, is like optimizing around a metric really hard? And that's the problem.

[0:08:07.6] DA: Yeah, I'm not going to argue with that. We actually have — Since I came to OpenAI, we got kind of an even more vivid example of this that we got to occur in an actual AI system. We are kind of playing with training simple video games and we have this boat race game where you're trying to complete a course with a boat. The only kind of easy way we have

of measuring progress. Again, the only simple way we have of measuring progress is it gets these points when it knocks down these targets along the course.

Kind of naively looking at it, it's something like, "Well, it has to knock down these targets to complete the course. Okay, we're going to give it — Train reinforcement learning system. Give it points for knocking down the targets. Great. It will complete the course."

I just set it in motion, didn't do anything for 24 hours. Then when I came back and I looked at it, the thing was going around in circles, because it found this lagoon where it could just get the maximum possible density of points.

[0:09:04.8] SC: I remember seeing that example.

[0:09:06.0] DA: Of course you can say, "Well. I don't know. You get what you asked for. The game was just broken." But I think the difficulty is that the mapping between what we think we're likely to get when we set the training process in motion and what we actually get, it's very discontinuous. It's not mysterious or magical or anything. You do indeed get what you asked for, but worries me is this kind of very unpredictable mapping between here's what we think we're trying to get a system to do, and here is what this system actually ends up doing. In retrospect, of course it made sense that it did that, but perspective predictability is a property that I'd like our system to have and that I believe they currently don't have.

[0:09:45.1] SC: One of the first things that you did around this research was to publish a paper. Again, it's the same paper I'm referring to with — I think it was in partnership with Google and another organization I forget the name of. You essentially outlined a set of rules. Do you think of them as rules or goals or —

[0:10:03.5] DA: Kind of general research area. Just to make sure, there's kind of two kind of major papers that we did in the last about year and a quarter. The first one I did while I was still at Google and it was done in collaboration with Google, OpenAI, Stanford Berkley, and this was kind of this agenda paper that kind of outlined various directions for AI safety.

Then more recently, about I think it was — What? Three months ago now. We published this paper called Learning From Human Preferences, which was kind of our research kind of attacking some subset of those problems. There's kind of like that. That paper was from a year ago was kind of the agenda doc that laid out.

[0:10:42.7] SC: That's the one I'm referring to. We'll dig into the next one in a sec.

[0:10:45.5] DA: Okay.

[0:10:46.4] SC: Maybe you can take a second to kind of — If you remember, like reading through those points in the agenda, because they're almost like — In some way, is it fair to think of them as like as [inaudible 0:10:57.3] new laws for the neural network age or something like that?

[0:11:00.2] DA: A little bit. Yeah. They're in a sense trying to solve the same problem, although it's less things a specific machine should do as kind of areas of research where to address particular types of problematic behavior that could arise.

There were kind of five research areas that we talked about and they were kind of divided into two topics similar to this kind of robustness versus value alignment thing. The kind of the schema we used was you have a machine learning system. You want it to do something and something goes wrong and it does something other than what you intended it to do. Where did things go wrong? Things go wrong because you have the wrong objective function and you optimize it really hard, or things could go wrong because you had the right objective function, but something about the way you trained it went wrong. That would be like the self-driving car that's put in a new environment or the kind of robot helicopter that is trying out behaviors and then destroys itself and it wasn't in the algorithm.

On the first one, the problems we talked about — The first one was reward hacking. This is kind of the thing we demonstrated in the boat race, where you have this measure, you optimize it really hard and you get the wrong thing. The second on in that area that we talked about was what we call negative side effects. The idea is the world is really big. Basically, it's a big similar

to reward hacking. Basically, any simple objective that I can come up with probably refers to very small set of things, right?

If I ask you to move this chair, I'm implicitly not referring to just every other thing under the sun that could be in the world.

[0:12:36.0] SC: Without breaking that window.

[0:12:36.6] DA: Yeah. By default I'm kind of like not specifying all of these commonsense stuff. You might say side effects are really bad by default in AI systems. That was kind of the second problem in that category. Those two are kind of like broad reasons why I picked out an objective function, seemed innocent, seemed good and something went very wrong with it.

The third problem which is kind of between the categories is this thing we called scalable super vision, which we dealt with that some in the later paper, which is even if I know — If I were there for everything an AI system did, even if I could coach it and tell it do the right thing, if I could supervise it every decision that it made. Make sure it never got out of my sight and never did anything. I don't really have the Bandwidth to do that. How do I handle situations where I kind of know what I want the AI system to do, but I have very limited — It's not feasible for me to have more than limited interaction with it. How do I handle that? That situation.

Those were the three on the kind of like how do I get the right objective function, and then I have the right objective function. How do I make sure something bad something happen? The two problems were safe exploration. Safe exploration is the idea that I have some objective function, like fly a helicopter. Even if that's right, even if I've set my system up in a way so that if I give it enough time, learns how to fly the helicopter right.

In the real world, if I have a robot helicopter and it crashes, maybe it breaks and it is never able to fly again. I have this kind of —

[0:14:07.1] SC: Can get expensive at the very least.

[0:14:09.4] DA: Yeah. You have what's called an asymptotic guarantee that you eventually get the right behavior, but it doesn't help you if you die before you get to the asymptotic limits. That's an area that has gotten some attention in machine learning already, but I think as with many things, kind of neural nets and kind of very powerful policies and new tasks have just come along in the last three or four years. So the safe exploration literature is just kind of catching up to this, and so one of the things we're saying in the paper was there's an urgency to making sure that these safety, the work on safety issues catches up with the work in other areas that's kind of hurdling ahead.

We've kind of already seen it. One example I give is at OpenAI we have this tool called Universe, we're using it for a while, that lest you basically have a reinforcement learning system that connects to the web and has, as its kind of actions and abilities, see in the screen, moving the cursor and clicking on anything.

Reinforcement learning systems, when you initialize them, tend to explore randomly. The first time I ever trained a reinforcement learning system on Universe, it immediately opened up Chrome, right clicked, open the Chrome developer tool's kit, changed the code in there, closed it, crashed Chrome and caused some kind of segmentation fault on my computer.

[0:15:31.7] SC: That's amazing.

[0:15:32.6] DA: The first time — Just random behavior. If your environment if complicated enough, you really, really have to think about these issues. I could say the same thing about — Google has done work on — I think it's well-known now, optimizing datacenters and using reinforcement learning to optimize datacenter energy usage. Of course, there are some knobs you could turn there that would break the datacenters. That is an issue that I think they've had to think about as well.

[0:15:58.5] SC: Right.

[0:15:59.6] DA: Kind of the last issue was this distributional shift thing which I've kind of alluded to a number of times and has to do with the environment and your self-car driving car changes or you train your speech system on one accent and you change it to another accent. One

situation which it came up was the infamous Goggle gorilla incident. Google photos tagged some African-American individuals as gorillas and it was kind of a problem with the training set, that the training set have been weighted towards Caucasian individuals, and so it got confused. It had no idea what racism was about or any of these things. Failures have kind of already happened and how can we be better about machine learning systems kind of knowing what they don't know. That's kind of the overview of the problems.

[0:16:44.8] SC: So the second paper is one that's kind of diving into kind of the first few of these problems.

[0:16:52.3] DA: Yeah, a couple of the first problems. As with most things, we wrote this kind of grand agenda and then we're like, "Oh! We have so many ideas for how we could work on any of these. Which should we start on?"

The thing we eventually settled on was kind of the reward hacking and scalable supervision stuff. Having the wrong objective function. With that, what we wanted to think about was, "Well, if you're trying to learn an objective function that's in line with what humans would want, then you should probably learn that objective function straight from humans, because if you have some kind of hard to place aesthetic thing that it's hard to encode into a system, instead of trying to encode it, you kind of let the human be the teacher, right?

In some sense, whenever we do supervised learning, we're doing that a little bit. A human has to label all the images to tell us, "Well, this is what a duck is. This is what an ostrich is." But it hasn't been done very much in the setting of reinforcement learning and kind of that's the setting that I think has both the most promise and that we should worry the most about.

[0:17:54.4] SC: Are you saying letting the human be the teacher necessarily by observation or by the human communicating a set of rules?

[0:18:02.8] DA: By the human communicating. I'll kind of explain a little bit how it works and how we set it up in our paper. For those who don't know, the usual set up with reinforcement learning is you have kind of an agent that's interacting in an intertwined way with an environment. You have some notion of reward. In the deep mind system that played Go which

did you win or did you lose in Atari? It's the score. Because you're training this system by trial and error for many millions of iterations through, it's actually very important in the way we currently train the system to have a kind of a programmatically, a valuable reward function.

With Go, if I had to have a human look at the end of every game and say, "Did you win or did you lose?" It would be very hard. It's very important the AI can just write a little script that says, "Yup. I have more territory than you. Therefore, I won. You have more territory than me. Therefore, you won."

In general, these systems are trained without much human intervention, because there is a time for it. What that means is that the goals have to be something you can write in a simple program. The way we changed this was what if you have a human every once in a while give some feedback on what the right goals are and whether the system is behaving in a way that it should behave.

The idea is with my reinforcement learning, I take out the reward function and I replace it with like a model of what the right thing to do is that is trained from a human. Concretely, the way it works, is I have my reinforcement learning system. It starts out by acting randomly as all reinforcement learning systems do, and then every once in a while, I take two clips of its behavior, just two randomly selected clips, give to a human, and the human takes a look at them and says, "Okay. This is a little more like what I want than that." The human has in mind some behavior that it wants the AI system to engage in. At the beginning, neither of the two video clips is going to be particularly good, but the human is like, "Yeah. This is a little more like what I want." Then the AI system has kind of a reward predictor that it trains that that model is the human choices. It tries to come up with a reward that's consistent with what the human chooses. It goes off an optimizes that for a while, then it presents the human with two more video clips and the human says, "This is better than that." Updates the system's model and then the system goes off and plays around in the environment and tries to achieve that goal better.

Instead of kind of careening off optimizing a particular goal really hard, you have this interplay where the system optimizes a goal, comes to the human and says, "Am I going in the right direction?" The human reinforces it or corrects it and then it optimizes some more.

Via this interplay, you can make sure that you're kind of gradually training a system that stays in the direction the human wants it go, while at the same time every time the human interacts, they're kind of imparting a little piece of what they have in mind to the AI system.

[0:20:59.2] SC: Can I jump in with a couple of questions here?

[0:20:59.8] DA: Yeah.

[0:21:00.7] SC: If the only source of data for the system to optimize around is the input is getting from the human, what it's doing between the times that it's getting input from the human, and like does it work if you just compress that and take it out?

[0:21:15.6] DA: Yeah. No, it actually doesn't, because if you kind of think about — Maybe it's good to have an example. Let's say you have a video game where you're kind of trying to shoot spaceships and keep them from shooting you. There're two different parts to it. There's understanding that the goal of the game is to shoot the spaceships and not be shot by enemy spaceships. There's the actual mechanical dexterity to find the spaceships, shoot them and avoid them.

[0:21:41.8] SC: It can instead learn how to do that stuff.

[0:21:43.5] DA: Yeah. Basically, what the system needs to do is it needs to figure out from you what its goal should be, namely; shooting the spaceships. As it figures out that goal, then it needs to, on its own, practice in the environment in order to learn how to achieve that goal.

In practice it's a little bit of an interplay. First, it understands that it should be avoiding these shots that are coming down. Then it kind of goes off and optimize that. Then it gives you some trajectories where it avoids the shots, but doesn't actually score any points and you're like, "Oh! This isn't any good." Then every once in a while it accidentally scores a point and you're like, "Yeah! You should do more of that." Then it kind of figures out, "Oh! I need to not just avoid the shots. I need to actually shoot the enemies." Then it takes some additional time to figure out how to do that. It's an interplay between the system learning what you want and the system learning how to achieve what you want. That's why there's that interplay, so you only ever see

about — In our paper it was about .1% of what the system actually does. We trained it on tasks on Atari. Atari is a common benchmark. Atari games are a common benchmark for reinforcement learning. We trained it on some Atari games, and the human only had to see about .1% of what the actual agent saw, which is good, because the agent sees days of experience or so. We don't have time to have humans see all of that.

[0:23:05.4] SC: Okay. What is the human evaluating — In a simple game, like a spaceship game you described, I'm imaging what is the human evaluating the two frames on? Is it just looking at the score? Is it a proxy for some way for it to detect the score or is it more than that?

[0:23:20.2] DA: We blank out the score, because it's a little bit of a confounder, but we just say to a human, "We're kind of, in some ways, trying to develop a pipeline such that you could take a task and actually form it out to humans who understand the task."

We blank out the score and we gave it to some contractors and we said, "This is a game where you're trying to shoot enemy spaceships and avoid getting shot by enemy spaceships." "You're going to see two video clips and click on the video clip that you think does the better job of that partciu8lar goal."

[0:23:49.4] SC: They're video, they're not still.

[0:23:51.9] DA: Yeah, they're video clips that are a couple of seconds long. So for many tasks, we're able to get a good sense of that. We're also able to do tasks where there's no — For Atari, there is a score that you could learn from, and our point was that you don't need it. You can learn without it. There are also tasks particularly kind of tasks that simulated robot animals do, where you want them to perform some trick that it's really hard to describe mathematically, but that a human can recognize. We taught kind of like little simulated robot walkers to do kind of backflips and frontflips. We taught like walkers with two legs to kind of like balance on their hind legs or do ballerina moves or things like that.

The exciting thing is, usually if you want to train something with reinforcement learning, you have to say, "Okay. What's the behavior I want? How do I write a mathematical function that assesses whether that behavior was right or not?" Whereas here it's just, "Okay. What's the

behavior I want? Let me look at some video clips and try and reinforcement that behavior." Which I think requires more human labor, but particularly if we form it out, it allows you to do a much wider variety of tasks.

[0:25:01.5] SC: Have you looked at — It sounds like you're pulling two random clips. Have you explored like selecting the clips based on maximal distance or information or something?

[0:25:12.1] DA: Yup. No, that's a really great question. We actually did include that in the paper and it helped a little. Instead of having kind of a single predictor of the human preferences from the data, one of the things we tried was having an ensemble of three predictors that are trained on subsets of the data. This is kind of common statistical validation technique.

Then you look at cases where different predictors disagree with each other about which clip they think the human would think was better. That disagreement is a good proxy for, "Oh! This is a hard case. This is a hard case to figure out."

Then we kind of mined for hard cases and preferentially present the human with hard cases instead of easy cases. In the spaceship example, it'd be like, you've really — Let's say the agent has really figured out that if you got shot by the ship, it's really bad. But then there are some intermediate cases where the ship's laser is shooting at you and it almost hits you, versus the one where it actually does hit you and like maybe you're predictor hasn't quite figured out the difference between those then. And so you really want to show those to the humans. The human can disambiguate.

We found that that in deed did speed things up on more tasks than it slowed things down. It wasn't a huge improvement, and we're actually looking for kind of more intelligent ways of doing this, because I mean that's really getting into having systems be active and ask us questions about the things they're confused about instead of receiving passive information about what we think they're confused about.

I think in the future, that's going to be a big part of it, that we want such teaching to be a dialogue between the machine and us just like it would be a dialogue between the teacher and the student or parents and the child or something like that.

[0:26:55.4] SC: The other interesting thing that jumped out at me and you describing that last example is like the ambiguity from the human's perspective. If the laser just misses the ship, is that like, "Oh! Bad" Because the ship was too close to the laser, "Oh! That's awesome." That ship was able to like avoid the laser. It was really close. Have you characterized that part of the problem at all?

[0:27:17.4] DA: Yeah. I think there we some kind of technical issues we ran into which is how do you define goal of the task? Another example, like the example you gave, is what if your spaceship just in a particular few second clip doesn't encounter any enemy spaceships at all? In other words, doesn't see anything. There's nothing it has to do. Is that good behavior, because it did what it was supposed to do, which is nothing? Or is it bad behavior because it's worse than a clip where it actually shot the spaceship?

In the formal reward formulism, you're basically supposed to say that that's worse behavior, because it didn't get the reward as supposed to the behavior where there was a ship and did get the reward. Of course, if I'm a human thinking about it, it's kind of ambiguous. We tried to give the instructions in such a way as to kind of resolve those ambiguities. We wanted to focus on, "Can we incentivize this behavior with clear one or two sentence instructions?"

The technical description of it is there's a difference between the reward, which is the immediate reward you get, the return, the advantage, which is how good a particular action is. Which of these technical concepts and reinforcement learning do we actually want humans to reinforce? There's a question of, "How easy is if for humans to understand these different settings, versus how much does it help the algorithm?" These are all kind of things that like have been explored a little bit in the literature, but, again, never with the level of difficulty of tasks. This is kind of one of the areas that we want to explore more and we consider really unexplored. This was really just the first paper and all these are like really great questions that we kind of ran into when thought about some, but you definitely haven't fully answered.

[0:29:02.8] SC: Is the plan to — Like you kind of peeled of two of those, is the plan to go even deeper on those two or to peel off another one?

[0:29:10.4] DA: Yeah. I think this direction we're pretty excited about, and so I think we're going to do a lot of things that kind of follow up this human preference learning. We see applications to robotics task, like tying particular types of knots in a rope. It's another task where the reward function is very hard to describe and people try and specify it. I think that may be the weak link. Maybe we can do things like that. Things that physical robots do in the world, things like dialogue system. Microsoft made this dialogue system called Tay. It ended up kind of spewing racist nonsense.

Whether a comment is offensive or not is precisely the kind of high-level aesthetic concept that you can't learn really well with a simple objective function. Could you train a restricted dialogue system to understand what racist or sexist or offensive comments are and never make them? Some of that is simple. It's just not using certain words, but the concept of something being offensive is also more subtle. I can say something that doesn't use any offensive words, but the content of what I'm saying might still be offensive. Can we train a system to understand those distinctions?

Actually, even things like safe exploration, what it means for exploration to be safe. Maybe that's something that we could learn through human preference learning.

[0:30:31.3] SC: Can you give an example of that?

[0:30:32.8] DA: There's been a lot of work on kind of learning to learn, and so you could imagine kind of learning to learn safely. If I'm a human and I am learning to play an Atari game, right? You could think of copying what the human does, but you could also think of copying the human's process of exploring the Atari game. There are different ways to explore it. I can explore recklessly, doing a lot of things with trial and error, or I can explore very carefully, making sure that nothing bad ever happens. For instance, if I'm a human remote controlling a helicopter, I'm not just going to try random things. I'm going to kind of gingerly try a few things at first that you don't make the helicopter crash, particularly if I think that it will break if it crashes. Then I'll kind of gradually get more bold. Can we train machine learning systems to learn in the same safe way that humans learn and could we do that by giving them feedback or examples or demonstrations of how humans learn? Then can we solve that problem?

That way we've also thought about kind of a richer forms of human feedbacks. Right now it's just like is left of better or right is better? Sometimes when I'm training one of those systems, I'm just like, "I want to tell you that that thing you did is really, really good." All I can do is just give that one left click, but really, really, this was the great thing you did and I —

[0:31:55.4] SC: When you're giving the example of the model's doing backflips and things like that, I was thinking like you need the Olympic panel with the 10s and the 9s and the 8.5s and all that.

[0:32:02.9] DA: Having a scaler dial is one thing. Ultimately, I just like to be able to give linguistic feedback. I'd like to say on this one, "Nice job. Do that again," or "Nice job, but go a little to the right." "Nice job, but that's only the first part of the move. After you do this, you need to do a flip."

Currently, our natural language processing isn't really up to that task, but that's kind of long term goal we'd work towards, and the real long term vision would be you have an AI system that is persistent in the world is doing things on your behalf and that you make sure that it always does the right things instead of the wrong thing by this continuous dialogue between you and it where you teach it what you want it to do and it gives you information and executes the things that it understands that you want it to do.

[0:32:51.3] SC: We've been talking mostly about explicit feedback. Are you also thinking about like the explicit versus implicit feedback might play into this?

[0:32:59.4] DA: What do you mean by implicit feedback?

[0:33:01.3] SC: For example, maybe the Tay example. In all of the examples we've been talking about, you are telling the system, "This is good. This is bad." We started with kind of this binary good/bad and we talked about scaler, continuous degrees of goodness and badness, but it's still like you're telling it explicitly and I guess I'm wondering if there are ways to either — Maybe I'm thinking also of like some of the emotional intelligence, like pick up from your reaction.

[0:33:35.8] DA: Yeah. Someone reacts in a way where they don't like what I'm doing, but they're not going to explicitly say, "That was a bad thing to do."

[0:33:42.0] SC: I guess at some point it's all numbers feeding into some models somewhere.

[0:33:46.3] DA: yeah. I think the natural language feedback is a little bit — It's kind of starting to get it, where I mean if people have kind of — People have ways of like lightly disagreeing or politely giving negative feedback and those are patterns that a machine learning system could, in principle, pick up on it the same time. I think in the end if we ever want something like that to happen, we're going to have to go kind of beyond just receiving feedback from humans to actually having models of humans. A human says something and the AI system says, "Okay. Here is my picture of the human and here are my hypothesis about why the human would say that particular thing." One of them doesn't want me to do this thing. It wants me to do something else.

That gets to kind of modeling of other minds. Theory of mind is something we talk about a lot in neuroscience. This is one of these things that I think AI systems will be able to get AI systems to do that once they have some kind of rudimentary theory of mind.

[0:34:46.5] SC: Even you saying that kind of makes me think of all the complexities, like the model trained on the Brit could not be used to interact with an American.

[0:34:53.4] DA: Yup. On the other hand, simple animals have theory of mind. If you have pets, dogs can do this, even mice can do this. They have very complicated pictures of how the humans around them work. They need to in order to survive. They need to know when humans are going to feed them. They need to know — Humans are going to be going to be unhappy with them. I think there are some hope of building systems that have some understanding of that short of full human level intelligence.

[0:35:21.8] SC: Yeah. Cool. Anything else that you guys are working on that you wanted to talk about?

[0:35:28.0] DA: Yeah. I think we're going in several directions with kind of the human feedback stuff. We have kind of a small team so far, but we're growing, so there's a lot of kind of projects starting that are either offshoots of this or kind of different directions on safety, but they're all pretty early, so there's not so much to say yet, but I'm hoping that in a few months we'll kind of have another batch of results that are kind of follow up on the human feedback stuff and also kind of new directions it's taking.

[0:35:57.9] SC: Nice. There's a website that's like 10,000 Hours or something like that that looks at these — I think looks pretty broadly at what they think are like career opportunities.

[0:36:10.4] DA: This is 80,000 Hours. I actually did a podcast with them.

[0:36:13.4] SC: I like underestimated it definitely.

[0:36:16.5] DA: 80,000 Hours is the length of a career, I think. I think that's why they chose that number. Yeah.

[0:36:22.1] SC: Their top career or their top opportunity is AI safety.

[0:36:26.6] DA: Yeah. I don't have any opinions on like — I would never claim that what I'm doing is like the most impactful thing you could possibly do. I did a podcast with them. I would say instead that the world has a lot of problems and a lot of issues, unfortunately like multiple incredibly serious issues.

One thing I think is true about AI safety is that if I look ahead to the next two decades or so, I believe it could be a really serious issue and I believe that not that many people are thinking about it seriously in the sense of doing actual technical work on it. I think for long time, it was dismissed maybe with some justification as kind of a crackpotty thing. I think that is starting to change.

I actually see opportunity there that there's this important problem that hasn't really been thought about in a careful enough way yet. To me, that creates an opportunity to work on a problem that's very important to the world that there are a giant number of smart people already

kind of descending on it. Because there's been kind of less thought about it, maybe by acting early we can actually come up with real solutions. To me, that's kind of my rationale for feeling that it's a high-impact thing for me to work on. That it seems important and it seems like there aren't already 100,000 people working on it.

[0:37:52.3] SC: Awesome. Dario, thank you very much for coming on the show.

[0:37:55.0] DA: Yeah. Thanks again for inviting me.

[END OF INTERVIEW]

[0:38:00.2] SC: All right everyone, that's our show for today. Thanks so much for listening and for your continued feedback and support. For more information on Dario or any of the topics covered in this episode, head on over to twimlai.com/talk/75.

To follow along with our OpenAI series, visit twimlai.com/openai. Of course, you can send along your feedback or questions via Twitter to @twimlai or @samcharrington or leave a comment right on the show notes page.

Thanks once again to NVIDIA for their support of this series. To learn more about what they're doing at NIPS, visit twimlai.com/nvidia. Of course, thanks once again to you for listening, and catch you next time.

[END]