Interview with David Duvenaud ("Interviewee") by Vael Gates ("VG"), 2/2/22

For context on this interview, see https://ai-risk-discussions.org/interviews.

—

VG (00:00):

Great, alright. So I have four sections of questions on perceptions of AI. So the first ones are just, this is my softball section, which is work motivations. So can you tell me what area you work on in AI in a few sentences, and how you came to work on this specific topic, which [is a specific subarea of deep learning research.]

Interviewee (00:00):

So the earliest I can really remember is finding Jürgen Schmidhuber's webpage. And I think it's a really underrated resource because he's explicitly making the case that things like beauty and humor, and creativity and surprise have—or can have—meaningful technical definitions, and are not something that's disconnected from the rest of theory and math and cognition, rather they're just parts of it. And I think a lot of people have made the case that emotions in general are not like some separate thing, that it's not like there's reason and then emotions, but actually it's all part of this strategy for making actions. [...]

Then I guess the biggest development was reading E.T. Jaynes's *Probability Theory: The Logic of Science*, which is like *Atlas Shrugged* for Bayesian ideology, I guess. It made it sound like there's this simple unified way to handle all of uncertainty, reasoning, inference, and it's being kind of unfairly, needlessly sidelined. I think he was right, and I think he lost over some of the issues a little bit, but for the most part, he was right. And he's also been vindicated I think in the last 40 years or 30 years since that book was written. Yeah. [...]

And then reluctantly started working on deep learning, because it was clearly just such an effective set of tools, even though they were intellectually unsatisfying. And, yeah, I guess the other major shift was just realizing that having a principled reason to use a method, it's just a really easy way to fool yourself into working on something, for reasons other than that it's a good idea. So, yeah, I think there's a lot of people in academia with pretty flimsy stories or reasons for working on the tools or niche that they're in. You have to find a niche. Most niches are not very interesting or important. And so you have to come up with a story for "Oh, but everyone else is going to be, well, everyone else is doing things for the wrong reasons." I have this principle by which my method is going to, in the long run, win. So, anyway, I mean, I might still be doing that.

Anyway, since then, I've been working on problems with machine learning. [Interviewee discusses a more specific research area.] And my current research agenda, I think, will result in better world-scale simulations, like weather simulations and stuff like that. But if I were trying my hardest to make AGI, I would not be focusing on this particular area. And part of the reason I'm in this area now is because it's a really fun niche, with lots of satisfying technical developments that are within my reach, and doesn't require giant engineering efforts on the scale of DeepMind and OpenAI. And so it's almost a lifestyle issue, where I like thinking about an idea with a few people, making proof of concept, getting it to work, and then spreading those ideas, and having other people implement them. But, again, from a results point of view, I should be forsaking my family, getting hundreds of researchers joining. Or if it's working with some larger organization, and running massive experiments all day, every day.

VG (02:49):

So my next question was, what motivates you in your work? So it sounds like you have this plan for what you would do if you wanted to build AGI, and you're not doing that. And, instead, you're not forsaking your family, and doing elegant things, and sharing information and doing technical problems.

Interviewee (03:08):

So, obviously, I don't think our motivations are very clear to us. [...] So I think status or something would probably be the number one. But it's hard for me to even think about that clearly. But everyone keeps asking me, "Why are you in academia, and not industry?" And I think it's because it's so fun to have interactions like this, where someone's asking your opinion about the world or whatever. Right? And I really like being a bit of a public figure, and not feeling like I'm beholden to some larger explicit hierarchy above me, even though we obviously are all just enmeshed in some horrible set of obligations. Not horrible, but just society anyway. So, anyway, I guess the other thing is it really bothers me not to be able to understand things, and just have to gloss over, and guess "Is this field legitimate or not, based on some correlates and features?" I really hate doing that. It's very unsatisfying. And the kind of people that I get to hang around is another huge motivator. And that's actually probably why I've been a hanger-on to the AI safety world. It's such a surprisingly awesome group of people. And I mean that literally. Some of the people, I've just been like, "Wow, like this really upgrades my expectations for what a human life can look like." And that's only happened to me a couple of times in my life. And it's been a huge deal every time it happened. And I just really want to be able to spend time with these people and be part of this project.

VG (04:48):Just because they have very insightful thoughts, or what do you mean?

Interviewee (04:56):

So one is living deliberately, and thinking hard about "What kind of life am I making for myself if I take these actions?" And the usual [...] practice of trying really hard to actually answer questions, and not just let yourself euphemistically say, "Oh, what will be your que sera, sera?" Or just find excuses to do the thing that you wanted to do anyways. There's a lot of ways in which this doesn't work, and it's not compatible with human relationships or motivations. And I've seen people wreck their lives by deciding that the thing that they want is the thing that's socially acceptable to want, and then trying to work backwards from there. I think we have to accept shitty-sounding motivations in our day-to-day interactions, which is a very awkward thing to do. I don't think people have really figured out how to do that very well. And I certainly don't push for that hard in my own personal life. [Interviewee talks about a

personal experience.] But at least in our own heads, it's a question of like, "Do you want to accurately predict your own future or not?" I guess. And there's trade offs. So people being extremely honest with themselves is really the main thing that impressed me. And that's all.

VG (06:25):

Attitudinal thing. That makes sense. All right. So now we're going to go ahead. That's the end of section one, which is work motivations. Now it's hopes and worries about AI, now and future. So the most general higher-level question is, what are you most excited about in AI? And what are you most worried about? So benefits, risks, excited, not excited, etc.

Interviewee (06:47):

So, as for benefits, I think it's just the normal ones we have, like, "Oh, cure disease." I guess I would also call myself a transhumanist, and say, "Oh, and I'd love to be part of some planet sized computer brain someday." And I think the human capacity for joy and all the positive aspects, we're only experiencing a tiny fraction of that. So it's hard for me to really concretely imagine that though. And I don't think human values are very meaningful in that regime anyways. And then I guess my worries are pretty mainstream among AI safety people, which is that there could be some runaway event of more powerful machines, or machines making themselves more powerful. I guess the way I differ from probably the main three narratives of Yudkowski would be to say, "I'm also just worried about our normal institutions, of doing the same thing." And I think we're, oh shit, well on the way to ending up in some permanent North Korea-like dystopia, whether we make AI or not. And there's a very simple argument, just from the point of view of dynamical systems, which is that, in the long run, they're going to end up in one basin of attraction or another. And so I guess I worry that our two choices are: I think Robin Hanson talks about the Hardscrapple Frontier. We're just going to build a bunch of von Neumann probes and spend all of our energy colonizing the universe. And that's maybe not going to be very interesting from a human values point of view. But at least there'll be fun engineering challenges and coordination problems that we'll have to overcome, and interact with whatever civilizations we meet, or we'll just end up in some horrible stable state here on Earth, where we've deliberately constrained our own growth. And that sounds horrible, pretty much no matter which way I cut it. And, yeah, I guess I just want to say, I sometimes wonder why AI safety people don't talk more about runaway governments or social movements or things like that. They're already pretty much as dangerous as they need to be to permanently ruin human civilization. Almost happened in the USSR. Think of the Cultural Revolution, but never-ending. It just happens on a regular basis. And the only thing that saves us is that these governments are incompetent, so eventually things break down, and there's disorder. And then you end up with local pockets of humans flourishing amongst the sea of misery. That's my answer.

VG (09:47):

So you are worried about dystopia happening with the current government systems. And the reason it hasn't happened yet is they're incompetent. And AI's going to add the incompetence? Or can this happen, like you said, without any AI?

Interviewee (09:59):

Yeah, exactly. So I think North Korea is a perfect example. It's surprisingly stable and robust. It's a 70-year-old regime now or something. So I think it can happen already without AI. And then I do think that AI is going to just make control easier. [Shares personal example of someone working on AI tools

that could be used for control, and not acknowledging the potential problems with this]. So I just feel like that just made me super pessimistic.

VG (11:14):

And that technology should be coming up pretty soon. I mean, we already have good surveillance technology, right?

Interviewee (11:19):

Well, I think Microsoft Word… Or maybe Google Docs just rolled out, like, "Oh, did you mean to say this more acceptable phrase sort of thing?" And that's the very mild end of the spectrum, right? That's not really what I'm worried about. The worry is just people being afraid to talk to each other honestly about whatever issue. And then the Overton window keeps shrinking and shrinking until we end up in an equilibrium where everyone hates it, but no one is willing or able to talk about it.

VG (11:47):

That makes sense. So I'm going to follow up on that higher level question, focus on future AI. So putting on our science fiction forecasting hat, say we're 50 plus years into the future, what does that future look like? So 50 plus can be as far out as you want. Are all jobs automated? Is society different? Are things mostly the same? Are we locked in?

Interviewee (12:12):

I predict, at the time, we'll have at least some AGI that is more reliable and useful than the average human worker. But it might be much more expensive. Might require a huge data center. So it still can't completely dominate humanity. But, even at that point, I think most humans will be economically marginalized, just because there's such a huge economic incentive to take humans out of the loop wherever we can. Because we're so unruly, and messy, and unreliable and stuff. And so I think I don't have a strong feeling for how well people will adapt to that. But I guess I feel like video games are going to be great. There's going to be even better drugs. Lots of sort of silence, it's not misery, but silence… People can start to stay in their rooms forever, enjoying whatever fruits of civilization, in a very cheap and cost-effective way. And then there will be some smaller faction of people who are still engaging in, I don't know, sort of human-like activities, like going out and doing things, and maybe raising families, or organizing and stuff like that. And then some small faction of Unabombers, tell people who would think this is horrible, and are trying to overthrow the system. But I think they'll be very effectively marginalized. So that's the end game I see, before we have actual stronger-than-human AI. And then, at that point, we might be able to uplift whatever fraction of these economically irrelevant people in some way. That's my best outcome, is they somehow become useful by being attached to some larger cognition system. But it would have to be in a fake way, I think. Because it's not like their personalities or life experiences would be really adding much to the equation, I think, in most instances, I would imagine. So that's my vague version of the future.

VG (14:26):

Oh man. It's not a very exciting future here. So 50 plus years.

Interviewee (14:31):

Well, I think. But I guess I'll say it's hard to imagine humans flourishing in detail, right? It's hard [inaudible]. Just human life, the good parts are just spread out and hard to describe. When people talk about hunter-gatherer societies, it doesn't sound all that great. It's like, "Oh yeah, everyone's just spending all their time hunting and gathering, or just telling stories around a campfire. That sounds crappy." But it doesn't have to be.

VG (14:59):

So is this a broadly pessimist or a broadly optimistic feature? On a scale one to 10, how good is this future 50 years out?

Interviewee (15:10):

Oh, I'd say that that's pretty optimistic. Because a lot of jobs today are super terrible. A lot of people are just trapped in misery, for reasons that I think we can fix within 50 years. And then I guess all I'm trying to say is I think the cool thing about our civilization right now is that it really does need people to run it. So people's contributions, however horrible they might be, like being a street cleaner or whatever, at least they are contributing to civilization, if that matters to them. I don't know. So maybe most people don't care about being as useful as I do. But I think people really want to feel useful. And I think they're right to feel that, if they stop being useful, that they will inevitably be somehow marginalized so that they have no power. And that's, again, what Dean [inaudible] wrote about, I think. My favorite analogy for AI dangers is just humans or, sorry, primates built something that was very slightly smarter than them, and very slightly unaligned. And now it does not matter what any monkey does. They have no say over the future of monkey society. They're just here at our pleasure, and are basically begging, until we decide to either build a nice zoo for them, or use the jungles that they live in for something else. And no funky monkey can ever give a speech that will change our minds, right? It just doesn't matter. So that's the boat that I see most of humanity being in, except those who have somehow uplifted themselves, or somehow merged with whatever computational soup is actually running things. I don't know. I guess you can't give me too much feedback on this, but I guess I would be surprised if this was really that different from what most people in the AI safety community think. But, to be honest, I haven't had that many long conversations about this either.

VG (17:10):

Happy to discuss at the end. And then my next question after that is a follow-up on risks in particular. So when you think about current day AI in particular, and then 50 plus year AI in particular, what do you think the major areas of concern are? And how concerning are they? So you've already done a little bit of this, but I think of looking at the comparisons.

Interviewee (17:34):

So I guess my main concern is AI interfacing with our current control and power structures, systematically misinforming everyone, which is so sad. One major update I made in the last few years is just how many people just still take the news seriously, even when it's clearly just super ideologically captured. I mean, the alternative is to waste hours every day on Twitter, trying to find the random accounts that happen to know what's actually going on. It's a very inefficient way of doing things. So, anyway, I guess I expect most people to willingly vote for their own destruction, because they've been convinced that it's ultimately going to be the right thing to do. And, again, this is not that different from the situation that I think we're already in today. I don't really have that many more specific dangers than that. I've read a lot of, I think, Paul Christiano, and Andrew Critch. [Talks about a private discussion about

the topic.] And they all have the same middle, which is that, even when people realize that this is dangerous and out of control, and that most of the decisions in our civilization are being made in this very opaque way, by some agents that we don't know if we can trust, it's too late to stop it. Because any resistance is going to be crushed either just economically, or it's going to be so devastating economically to turn off the total plug, that we can't do it. Even just people in the system that'll be so invested in the status quo, that if you start saying, "Hey, we need to go back to human control or something," people will marginalize you very effectively. So I think we have a pretty short window. I mean, again, we've already kind of passed it, right? Even in the West, it's just really hard to criticize most aspects of most government control, and really marginalized people now, especially that it's not necessary or sufficient to meet in person anymore to do politics. And so I think that that release valve is kind of going away. If I decided that the Internet was very effectively censored, I couldn't effectively organize a mass movement just by walking around in person anymore, I think.

VG (20:22):

And you don't expect this trend to reverse anytime soon, it sounds like?

Interviewee (20:27):

No.

VG (20:27):

No?

Interviewee (20:28):

No. I mean, I have hopes. I feel like a lot of people are becoming more skeptical in the last few years. That's one upside of COVID to me, is that I was very skeptical about a lot of our institutions before COVID, and now I feel these people saying, "What? Why are the public health people acting like this?" And I'm like, "Yes. Join me, join me." But ultimately, again, it's easy to criticize. And we can build some new institutions. But I think that they're going to have similar dynamics, at least over the long run.

VG (21:02):

Sort of unrelated question, but do you have timelines for AGI?

Interviewee (21:07):

I guess maybe let's say 10, well obviously, this is all uncertain and probabilistic. But let's say, something like five or 10 years from now, we start to have these expensive AIs that can do things reasonably well, compared to humans in some domains, for intellectual work, like lawyers or something like that. But because they're so… You can copy them and you can really focus development time on them. As long as you can make one that's pretty expert, then you can copy its output a lot. So that's when people like white collar workers will start to be economically marginalized. And then there'll probably be some pushback. The economic incentives are going to be so strong at that point, to just drive down the cost of compute, and running these models, that five or 10 years after that, there will just be so much compute being run that it will be comparable to the compute of the top 10,000 most influential humans or something like that. And then everything will change really quickly. But that's just vague.

VG (22:25):

Well, in 10 years, we'll have sort of AGI, but very expensive? And, then 20 years out, we'll have cheaper AGI?

Interviewee (22:38):

Yeah. And let's say that's the mode of a log normal. So probably maybe 10 or 20 years longer than that. I mean, I'm not happy. I'm not happy with that. I don't know.

VG (22:52):

Indeed. Section three is about external perceptions. So if you could change your colleagues' perceptions of AI, what attitudes or beliefs would you want them to have? So what beliefs do they currently have, and how would you want those to change?

Interviewee (23:09):

Yeah, generally, I guess the thing that sticks in my craw right now is the discussion of bias. And my own colleagues said, "Oh yeah. So I heard about this ProPublica article, on this COMPAS algorithm, and I..." And these people who actually use machine learning in their research. And they're like, "Oh, and I totally thought that it was a bunch of racist researchers who just put their thumb on the scale, and made a racist algorithm." And then, years later, they kind of learned, "Oh, actually, it's one of these things where there's a trade-off between being accurate and calibrated, versus putting your thumb on the scale to make the racial averages more similar." And that the ProPublica article was basically slamming them for not doing the second thing, instead of trying to be well calibrated. Although that particular article, the claims are more messy than that. But the point is, I think the article leaves this question deliberately blank. They never quite claim one way or the other. And then the reporting sometimes makes it clear, and then sometimes just makes the inflammatory claim explicitly. And thus the articles get spread around. And I understand that this is how journalism works, and this is how probably half the ideas in my head have been informed, but I found that very sad and frustrating. And not to say there's no issues there, but I wish people just understood how much they're being misled, and how weak and qualified the claims that make it to the headlines are, about AI bias basically. I mean, but that doesn't really matter. But I guess I feel like this is the vector by which control of AI becomes politicized. As people can say, "Well, but isn't it getting racist?" And I don't know. Our public discourse gets upgraded, I think people rightly realize that there is a power center here being developed. And it's almost always worth fighting over the power. So whether or not we solve this particular issue, it's going to be politicized one way or another. And we're just going to have a whole bunch of horrible, pointless, zero sum arguments. And it makes me sad. So, sorry, I think it's a concrete question. And then later around towards society again, and pessimistic predictions.

VG (25:54):

I also wanted to track back a bit, and say, how likely do you think it is that we're going to have a great future? Or do you think the problems of control are going to be solved, or of institutional epistemics?

Interviewee (26:07):

I'll say, from the point of view of whatever future thing takes over, I think they're going to be having a great time. And from the point of view of us today, I think we'll have a really hard time preserving our values. So it's just like our 100 year old ancestors would be shocked by how we live our lives or something, in one way or another. Like not being religious enough or something. And I can't say that they're wrong. Maybe, from their point of view, we might just be horrible degenerates. So, just based on

that having happened continuously throughout evolution, all the single cell organisms are like, "Oh, what the hell are these communistic, multicellular life doing?" No one is really happy about their descendants mutating into these new life forms, right? Unless you have a really weird specific set of values that says, "I don't really care what I turn into, it'll be fine." Which I think we do, to some extent, but not that much.

VG (27:08):

And these future people who are having a good time, are they going to be AIs mainly? Or are they going to be uploaded people?

Interviewee (27:16):

Yeah. At least, I'd say, they'll probably be living on computers. It's just a much better substrate for scalability. So whether they're simulations of humans, I don't think that really changes things all that much. Because we're not going to keep that many of our cells around if we simulate ourselves. It's just really expensive for not that much gain. Maybe it's fine, right? We'll say we'll keep around the important parts. We'll get rid of detailed hair simulation and poop simulation or something, right? And that won't be a big loss. But by most measures, it probably won't be something we would consider all that human. And there's definitely just going to be so much value stripped. Once you're uploaded, you're probably going to find it really easy to convince yourself to change a few more things that make your life way better from your own point of view, that your old self would've probably thought were a little bit weird, like merging with, I don't know, 1,000 other brains. I have no idea.

VG (28:14):

All right. Back to the original section. So if you could change your colleagues' perception, how about the public policymakers, and the media, are there things that should be happening right now?

Interviewee (28:30):

I don't really think so. [inaudible] said, "What should people have been doing in 1939 to prepare for nuclear weapons?" I really like how things are going actually, in terms of AI safety, which is that there's this weird little community of very serious, dedicated, impressive, thoughtful people. And there's lots of money sloshing around there, I think more than even we know how to spend right now. But I'm really afraid of growing the field much faster than we are. Statistics is a field that just became shitty, and then ossified and wasted everyone's time for like 50 years. And it's still around. And as a stats professor, I'll say, it's a relatively shitty field compared to computer science. And you can't undo that. You can't just find all the people who are hangers on, who care about stupid questions, and fire them. That's just never going to be feasible. So, anyway, I guess I feel like what's being done right now is exactly what I hope would be being done. And then actually I just noticed last week on Twitter, [inaudible] started putting AI safety in her sights, and saying, "This is a horrible field. These are basically for being consequentialists, and shut it down. This is icky and stuff." So I think there's a coming war. And I don't know what's going to happen. I'm really not clear how it's going to go. But there's no concrete actions that I can think to take right now. Maybe just getting a more under [inaudible] aware that this is a subfield. Because I do think that there are people who just don't get into it because it's so obscure right now. But I guess I'll say, at some point, it's probably going to become high-status. And that's the beginning of the end. I mean, again, I said I do things because of status. But I do think that fields are usually at their best right before they become high-status. And when they're starting to make traction. Okay, the amount you're motivated by the work and the outcomes, versus how much you're motivated by status is a matter of degree. And I guess I'll just say, in my vague understanding, fields usually work best when people are

more interested in the work and the outcomes than in the status. The status-oriented people still do good work for a while, but I think those are the ones who let the field drift. And they don't really care if we spend a lot of time answering pointless questions, or various parasites attached to the organization, or it's like governments dictate where those are. They're not going to fight hard to avoid those outcomes, is my guess. I don't really know that much about organizational dynamics.

VG (31:29):

What are the areas you're most excited about in AI safety, or the research you're most excited about?

Interviewee (31:37):

I mean, I do like this practical research that's been happening lately. I ultimately think it's doomed, in the sense that I don't think it's going to be able to make agents that are all that or as aligned as we'd like them to be.

VG (31:59):

What's the practical research, like examples?

Interviewee (32:00):

Yes, to give a concrete example, Anthropic has been working on papers, and opening eyes, saying, "Okay, let's take an actual language model, or reinforcement learning algorithm or whatever, and on a simulated task, try to get it to do something that, later, if we ask you to generalize it, does something slightly new." So the behaviors that the humans still think that it's a reasonable, aligned thing to do. And, yeah, it's the least unsatisfying approach that I can think of, because I don't know. I guess I feel like people have already identified a bunch of really hard impossibility results, about optimizing a reward function that you never get to observe. You're just never really sure that people aren't consistently somehow wrong, or lying, or unable to articulate something that is really important to them. And you still have to act on their behalf. And there's going to be days when the person is saying, "No. Hey, can you stop? This is not what I want." And you either take their word for it, and then never really do anything beyond the obvious things to make their lives better, and you probably end up wireheading them, if you go down that route, basically. Or, like a parent, you say, "No, for your own good, I'm going to make you upload yourself or whatever", and hope that you're doing the right thing. And so I just think there's no way around that. At least scanning training data of what people say that they want, it's also pretty unsatisfying for, I think, pretty obvious reasons. But I can't think of anything better.

VG (33:47):

So you're glad this field exists. There are approaches that are maybe currently doomed, but hopefully they'll figure it out?

Interviewee (33:56):

They're doomed, but they're doing about the best thing that I can imagine people doing. Okay. I guess I'll say there's a lot of fields where I'm like, "You're..." Well, not a lot. Some particular subfields that I'm familiar with, I've seen people who I think are very bright, spend their whole adult lives pursuing a direction that is wrong, for reasons that they should be cognizant of. They have all the tools to put it together, and see this is a dead end approach. And I don't see that happening in the AI safety field, which is amazing. Yeah.

VG (34:28):

All right. So you believe in the direction the field is going, the way they think, and how they're thinking about the problems.

Interviewee (34:35):

Yeah. There's no obvious incompetence, which is a surprisingly high bar.

VG (34:43):

Do you think they will develop less doomed things later on?

Interviewee (34:49):

Yeah. Well, the thing is that I don't think that there is a satisfying answer. So even if we said, "My utility function was like... We were all trying to optimize…" There's always going to be a way that I could look at a particular future and say, "This is horrifying," by some principle of mine. And so now the fact that there's like 7 billion people on Earth, some fraction of people are just going to find any possible future horrifying. And there might be some clever way that we can somehow divide reality or fake trick each other into thinking different things are happening. But, yeah, I guess I feel like there's an upper bound on how much even one person can agree on something being a good idea, in all senses of being consistent. So, again, I think we're probably going to be horrified by the future, and that's the best we can hope for.

VG (35:39):

All right. Being horrified by the future.

Interviewee (35:40):

Yeah. But there's degrees, right? If we're not personally being tortured, that's what I'm trying to avoid. Or just putting our planet into some sort of weird stasis, I guess. Obviously, that's still way better than being tortured anyway. I don't know. I guess your reaction makes me feel like I'm some sort of outlier here. There's a lot of people who spend a lot more time thinking about this than I have. And, as far as I can tell, I'm just coming to similar conclusions as them. But maybe I'm not. Maybe I'm an outlier without realizing it.

VG (36:24):

One of my questions was how much time have you spent thinking about each of these, I guess, in terms of hours?

Interviewee (36:35):

Yeah. So I've easily spent hundreds of hours thinking about AGI and distant future scenarios. And I have also spent probably hundreds of hours trying to do AGI research. [Interviewee names some research collaborators.]

VG (36:53):

AGI research?

Interviewee (36:56):

Yeah, probably on the order of 1,000 hours of my life so far.

VG (37:00):

And AGI research is to make AGI, or to do safety things?

Interviewee (37:05):

Yeah. I meant to say AI safety. Sorry. No, 1,000 sounds too high. I mean, let's say, again, 400 or 500.

VG (37:21):

Okay. Great. And I was about to say, have you taken any actions? Or would you take any actions in your work to address perceived risk? For me, it sounds like you have.

Interviewee (37:28):

Well, okay, so this is something I'm a little ashamed of personally, actually. [Talks about a personal experience.] I guess one thing I want to say is that what usually happens is that we identify some problems, and we realize that all we can really do is articulate this issue, and then demonstrate it in a toy world. And then that's the paper. And it feels very unsatisfying. Because once you've thought about it for a little while, it seems totally obvious. And pretty much every time this has happened, someone else has written the exact same paper, and it's actually been a valuable contribution to the field. So a concrete example would be, if you're trying to learn someone's preferences from their behavior, if they make a mistake because they're just dumb or distracted, you might be tempted to think that they actually wanted to crash their car, or they enjoy pain or something. But, of course, that's not true. Their bounded rationality or whatever, their cognitive limitations, as well as their actions, if you want to ask what they kind of actually would've done, if they hadn't screwed up, basically. And, again, it's a kind of trivial point. But I do think it's still worth writing the paper. So, in hindsight, I wish we had just gotten some of those out the door. So, anyway, the point is that that's some actions that I've taken that I gave up on halfway. So they're invisible. And I wish I hadn't given up on them halfway now. But, beyond that, that's about it. Yeah, I can't really point to any concrete contributions I've made to the field, or actions that I've taken that I expect would actually help [...].

VG (39:30):

When did you first hear about AI safety?

Interviewee (39:53):

[Interviewee mentions an early example of discussing the concept of AI safety.]

VG (40:21):

So considering your colleagues now at your institution, have other people heard about AI safety? Or do people talk about it?

Interviewee (40:33):

Oh, certainly. [Interviewee discusses some concrete examples.] So I guess I'll say it's definitely entering the mainstream. And, as it does, its focus is becoming more mainstream. And it's being deliberately conflated with making a robot nanny that won't hurt your kid, or a robot factory. A robot factory worker that won't cause injuries or something like that, which I think is probably bad overall. I think it's really

rare when civilization gets to actually spend resources on these weird questions, like AI safety. You can only really demonstrate their worth by argument. People should generally not trust themselves to evaluate arguments. So I guess I'm a little bit worried about… There being a big [inaudible] institute for AI safety, and it's all about self-driving cars or something like that. That won't really be a terrible outcome though. It doesn't really hurt anything that much. [Talks about a specific example where a field becoming more mainstream had undesirable consequences.]

VG (42:39):

So it's not a good thing that it's becoming more mainstream.

Interviewee (42:45):

Yeah, no. Yeah, but it's not terrible. I think that there's enough people who have drank the hard version of the AI safety Kool-Aid, that the money will be around forever. Just [mentions a funder] is enough, right? AI safety research isn't all that expensive.

VG (43:06):

All right. So I noticed we have a minute left. So is there anything else that you'd want to share, that I haven't asked yet?

Interviewee (43:30):

No, not really. I guess the biggest thing I'll say that I'm getting from this interview is me realizing how uncertain I am of how mainstream my own views are within the AI safety field. So I would love later to read your paper, to see what other people thought.

VG (44:22):

Yeah. [...] Well, I'm at Berkeley, and in the AI safety community. So I hear all the arguments, and reading all the merry discords, and all the fun stuff. So I don't think it's very dissimilar to what you believe, necessarily. Yeah. I think I'm a little bit less worried about current technologies and current institutions, though maybe I should be more.

Interviewee (44:43):

[There are] pretty good ones in the West right now, but I think they're just heading towards the failure modes that we've seen in the East, basically.

VG (44:50):

Yeah, that makes sense. Cool. All right. Well, thank you so much. Yeah. I'm generally curious about long-term risks, but I'm exploring what other people think about this, and what they think about the future, and current trends. So very much appreciate you taking the time.

Interviewee (45:07):

Oh, my pleasure. My pleasure. Good to see you.