Interview with cvgig, on 3/24/22

0:00:02.5 VG: Awesome. Alright. So my first question is, can you tell me about what area of AI you work on in a few sentences?

0:00:09.8 Interviewee: Yeah. So I'm what's technically called a computational neuroscientist [...]

0:00:44.0 VG: So these questions are like, AI questions, but feel free to like-- (Interviewee: "No, go ahead.") Okay, cool. Sounds good. Alright. What are you most excited about in AI and what are you most worried about? In other words, what are the biggest benefits or risks of AI?

0:00:55.9 Interviewee: Right. So in terms of benefits, I think that my answer might be a little bit divergent again, because I'm a computational neuroscientist. But I think that AI and the tools surrounding AI give us a huge amount of power to understand both the human brain, cognition itself, and more general phenomena in the world. I mean, you see AI used in physics and in other areas. I think that it is just a very powerful tool in general for building understanding. In terms of risks, I think that it's, again, by virtue of being a very powerful tool, also something that can be used for just a huge number of nefarious things like governmental surveillance, to name one, military targeting technology and things like that, that could be used to kill or harm or disenfranchise large numbers of people in an automated way.

0:02:04.2 VG: Awesome, makes sense. Yeah, and then focusing on future AI, putting on a science fiction forecasting hat, say we're 50-plus years into the future. So at least 50 years in the future, what does that future look like? This is not necessarily in terms of AI, but if AI is important, then include AI.

0:02:22.6 Interviewee: Yeah, so 50-plus years in the future. I always have trouble speculating with things like this. [chuckle] I think it'll be way harder than people tend to be willing to extrapolate. And also, I think that AI is not going to play as large of a role as someone might think. I think that... I don't know, I mean in much the same way, I think it'll just be the same news with a different veneer. So we'll have more powerful technology, we'll have artificial intelligence for self-driving cars and things like that. I think that the technologies that we have available will be radically changed, but I don't think that AI is really going to fundamentally change the way that people... Whether people are kind or cruel to one another, I guess. Yeah, is that a good answer? I don't know. [chuckle]

0:03:21.8 VG: I'm looking for your answer. So...

0:03:26.3 VG: Yes. 50 years in the future, you're like, it will be... Society will basically kind of be the same as it is today. There will be some different applications than exists currently.

0:03:36.4 Interviewee: Yeah, unless it's... It's perfectly possible society will utterly collapse, but I don't really think AI will be the reason for that. [chuckle] So, yeah, right.

0:03:47.5 VG: What are you most worried about?

0:03:50.9 Interviewee: In terms of societal collapse? I'd say climate change, pandemic or nuclear

war are much more likely. But I don't know, I'm not really betting on things having actually collapsed in 50 years. I hope they don't, yeah. [chuckle]

0:04:07.0 VG: Alright, I'm gonna go on a bit of a spiel. So people talk about the promise...

0:04:10.7 Interviewee: Yeah, yeah.

0:04:12.6 VG: [chuckle] Yeah, people talk about the promise of AI, by which they mean many things, but one of the things they may mean is whether... The thing that I'm referencing here is having a very generally capable system, such that you could have an AI that has the cognitive capacities that could replace all current day jobs, whether or not we choose to have those jobs replaced. And so I often think about this within the frame of like 2012, we had the deep learning revolution with AlexNet, and then 10 years later, here we are and we have systems like GPT-3, which have some weirdly emergent capabilities, like they can do some text generation and some language translation and some coding and some math.

0:04:42.7 VG: And one might expect that if we continue pouring all of the human effort that has been going into this, like we continue training a whole lot of young people, we continue pouring money in, and we have nations competing, we have corporations competing, that... And lots of talent, and if we see algorithmic improvements at the same rate we've seen, and if we see hardware improvements, like we see optical or quantum computing, then we might very well scale to very general systems, or we may not. So we might hit some sort of ceiling and need a paradigm shift. But my question is, regardless of how we get there, do you think we'll ever get very general systems like a CEO AI or a scientist AI? And if so, when?

0:05:20.6 Interviewee: Yeah, so I guess this is somewhat similar to my previous answer. There is definitely an exponential growth in AI capabilities right now, but the beginning of any form of saturating function is an exponential. I think that it is very unlikely that we are going to get a general AI with the technologies and approaches that we currently have. I think that it would require many steps of huge technological improvements before we reach that stage. And so things that you mentioned like quantum computing, or things like that.

0:06:00.1 Interviewee: But I think that fundamentally, even though we have made very large advances in tools like AlexNet, we tend to have very little understanding of how those tools actually work. And I think that those tools break down in very obvious places, once you push them beyond the box that they're currently used in. So, very straightforward image recognition technologies or language technologies. We don't really have very much in terms of embodied agents working with temporal data, for instance. I think that...

0:06:42.2 Interviewee: I essentially think that even though these tools are very, very successful in the limited domains that they operate in, that does not mean that they have scaled to a general AI. What was the second half of your question? It was like kind of, Given that we... Do you have it what it'll look like, or...

0:06:57.4 VG: Nah, it was actually just like, will we ever get these kind of general AIs, and if so, when? So...

0:07:03.2 Interviewee: Yeah, so I would essentially say that it's too far in the future for me to be

able to give a good estimate. I think that it's 50 plus years, yeah.

0:07:13.2 VG: 50 plus years. Are you thinking like a thousand years or you're thinking like a hundred years or?

0:07:19.7 Interviewee: I don't know. I mean, I hope that it's earlier than that. I like the idea of us being able to create such things, whether we would and how we would use them. I would not, [chuckle] I don't think I would want to see a CEO AI, [chuckle] but there are many forms of general artificial intelligence that could be very interesting and not all that different from an ordinary person. And so I would be perfectly happy to see something like that, but I just, you know, and I guess in some sense, my work is hopefully contributing to something along those lines, but I don't think that I could guess when it would be, yeah.

0:08:00.2 VG: Yeah. Some people think that we might actually get there via by just scaling, like the scaling hypothesis, scale our current deep learning system, more compute, more money, like more efficient, more like use of data, more efficiency in general, yeah. And do you think this is like basically misguided or something?

0:08:15.9 Interviewee: Yeah, let me take a moment to think about how to articulate that properly. I think... Yeah, you know, let me just take a moment. I think that when you hear people like, for instance, Elon Musk or something along these lines saying something like this, it reflects how a person who is attempting to get these things to come to pass and has a large amount of money would say something, right. It's like, what I'm doing is I'm pouring a large amount of money into this system and things keep on happening, so I'm happy with that. But I think that from my position of seeing how much work and effort goes into every single incremental advance that we see, I think that it's just, there are so many individual steps that need to be made and any one of them could go wrong and provide a, essentially a fundamental ceiling on the capabilities that we're able to reach with our current technologies. And so it just seems a little, a little hard to extrapolate that far in the future.

0:09:25.5 VG: Yeah. What kind of things do you think we'll need in order to have something like, you know, a multi-step planner can do social modeling, can model all of the things modeling it like that kind of level of general.

0:09:35.5 Interviewee: Yeah. So I think that one of the main things that has made vision technologies work extremely well is massive parallelization in training their algorithms. And I think that, what this reflects is the difficulty involved in training a large number... So essentially, when you train an algorithm like this, you have a large number of units in the brain like neurons or something like that, that all need to change their connections in order to become better at performing some task. And two things really tend to limit these types of algorithms, it's the size and quality of the data set that's being fed into the algorithm and just the amount of time that you are running the algorithm for. So it might take weeks to run a state-of-the-art algorithm and train it now. And you can get big advances by being able to train multiple units in parallel and things like that.

0:10:33.5 Interviewee: And so I think that the easiest way to get very large data sets and have everything run in parallel is with specialized hardware called, you know, people would call that wetware or neuromorphic computing or something along those lines. Which is currently very, very new and has not really, as far as I know, been used for anything particularly revolutionary up to this

point. You can correct me if I'm wrong on that. I would expect that you would have to have essentially embodied agents before you can get... in a system that is learning and perceiving at the same time before you could get general intelligence.

0:11:12.5 VG: Well, yeah, that's certainly very interesting to me. So, it's not... So people are like, "We definitely need hardware improvements." And I'm like, "Yup, current day systems are not very good at stuff. Sure, we need hardware improvements." And you're saying, are you saying we need to like branch sideways and do wetware-- these are like biological kind of substrates, or are they different types of hardware?

0:11:37.3 Interviewee: I guess different types of hardware is maybe the shorter term goal on something like that. Like you would expect circuits in which individual units of your circuit look a little bit like neurons and are capable to adapt their connections with one another, running things in parallel like that can save a lot of energy and allows you to kind of train your system in real time. So it seems like that has some potential, but it's such a new field that, this is when I, when I think about what time horizon you would need for something like this to occur, it seems like you would need significant technological improvements that I just don't know when they'll come.

0:12:20.4 VG: Yeah. So I haven't heard of this wetware concept. So like it's a physical substrate that like... It like creates, it creates new physical connections like neurons do or it just like, does, you know...

0:12:33.5 Interviewee: No, it doesn't create physical connections. You could just imagine this like... So, you know, computer systems have programs that they run in kind of an abstract way.

0:12:43.8 VG: Yep.

0:12:44.8 Interviewee: And the hardware itself is logic circuits that are performing some kind of function.

0:12:48.9 VG: Yep.

0:12:49.8 Interviewee: And neuromorphic computing is individual circuits in your computer have been specially designed to individually look like the functions that are used in neural networks. So you have... Basically, the circuit itself is a neural network, and because you don't have these extra layers of programming added in on top, you can run them continuously and have them work with much lower energy and stuff like that. It's just... It's limiting because they can't implement arbitrary programs, they can only do neural network functions, and so it's kind of like a specialized AI chip. People are working on developing that now... Yeah.

0:13:32.7 VG: Okay, cool, so this is one of the new hardware-like things down the line. Cool, that makes sense. Alright, so you'd like to see better hardware, probably you'd say that you'd probably need more data, or more efficient use of data. Presumably for this-- because the kind of continuous learning that humans do, you need to be able to have it acquire and process continuous streams of both image and text data at least. Yeah, what else is needed?

0:14:03.8 Interviewee: Oh, I think that... Yeah, more fundamentally than either of those things. It's just the fact that we don't understand what these algorithms are doing at all. And so we're... You can

train it, you can train an algorithm and say, "Okay, you know, it does what I want it to do, it performs well," and most machine learning techniques are not very good at actually interrogating what a neural network is actually doing when it's processing images. And there are many instances recently, I think the easiest example is adversarial networks, if you've heard of those?

0:14:41.8 VG: Mm-hmm.

0:14:42.2 Interviewee: I don't know what audience I'm supposed to be talking to in this interview.

0:14:46.4 VG: Yeah, just talk to me I think.

0:14:49.2 Interviewee: Okay, okay.

0:14:50.1 VG: I do know what adversarial... Yeah.

0:14:52.9 Interviewee: Okay, so, adversarial networks are... You perturb images in order to get your network to output very weird answers. And the ability of making a network do something like that, where you are able to change its responses in a way that's very different from the human visual system by artificial manipulations, makes me worried that these systems are not really doing what we think they're doing, and that not enough time has been invested in actually figuring out how to fix that, which is currently a very active area of research, and it's partly limited by the data sets that we've been showing our neural networks. But I think in general, there's been too much of an emphasis on getting short-term benefits in these systems, and not enough effort on actually understanding what they're learning and how they work.

0:15:43.5 VG: That makes sense. Do you think that the trend... So if we're at the point where people are deploying things that you don't understand very well, do you think that this trend will continue and we'll continue advancing forward without having this understanding, or do you think it would catch up or...

0:16:00.4 Interviewee: Yeah, well, I think it's reflective of the huge pragmatic influence that is going on in machine learning, which is essentially, corporations can make very large amounts of money by having incremental performance increases over their preferred competitors. And so, that's what's getting paid right now. And if you look at major conferences, the vast majority of papers are not probing the details of the networks that they're training, but are only showing how they compare it to competitors. They'll say, "Okay, mine does better, therefore, I did a good job," which is really not... It's a good way to get short-term benefits to perform, essentially, engineering functions, but once you hit a boundary in the capabilities of your system, you really need to have understanding in order to be able to be advanced further. And so I really think it's the funding structure, and the incentive structure for the scientists that's limiting advancement.

0:17:02.2 VG: That makes sense. Yeah, and again, I hear a lot of thoughts that the field is this way and they have their focus on benchmarks is maybe not... and incremental improvements in state-of-the-art is not necessarily very good for... especially for understanding. When I think about organizations like DeepMind or OpenAI, who're kind of exclusively or... explicitly aimed at trying to create very capable systems like AGI, they... I feel like they've gotten results that I wouldn't have expected them to get. It doesn't seem like you could should just be able to scale a model and then you get something that can do text generation that kind of passes the Turing Test in some ways, and

do some language translation, a whole bunch of things at once. And then we're further integrating with these foundational models, like the text and video and things. And I think that those people will, even if they don't understand their systems, will continue advancing and having unexpected progress. What do you think of that?

0:18:09.6 Interviewee: Yeah, I think it's possible. I think that DeepMind and OpenAI have basically had some undoubtedly, extremely impressive results, with things like AlphaGo, for instance. What's it called, AlphaStar, the one that plays StarCraft. There are lots of really interesting reinforcement learning examples for how they train their systems. Yeah, I think it just remains to be seen, essentially. It would be nice-- Well, maybe it wouldn't be nice, it would be interesting to see if you can just throw more at the system, throw more computing capabilities at problems, and see them end up being fixed, but I...

0:19:04.0 Interviewee: I'm just skeptical, I guess. It's not the type of work that I want to be doing, which is maybe biasing my response, and I don't think that we should be doing work that does not involve understanding for ethical reasons and advancing general intelligence. For reasons that I stated, that essentially, if you hit a wall you'll get very stuck. But yeah, you're totally right that there have had been some extremely, extremely impressive examples in terms of the capability capabilities of DeepMind. And, yeah, there's not too much to be said for me on that front.

0:19:46.8 VG: Yeah. So you said it would be interesting, you don't know if it would will be nice. Because one of the reasons that it maybe wouldn't be nice is that you said that there's ethical considerations. And then you also said there's this other thing; if you don't understand things then when you get stuck, you really get stuck though.

0:20:01.5 Interviewee: Yeah.

0:20:04.4 VG: Yeah, it seems right. I would kind of expect that if people really got stuck, they would start pouring effort into interpretability work for other types of things.

0:20:12.7 Interviewee: Right. You would certainly hope so. And I think that there has been some push in that direction, especially there's been a huge... I keep on coming back to the adversarial networks example, because there have actually been a huge number of studies trying to look at how adversarial examples work and how you can prevent systems from being targeted by adversarial attacks and things along those lines. Which is not quite interpretability, it's still kind of motivated by building secure, high performance systems. But I think that you're right, essentially, once you hit a wall, things come back to interpretability. And this is, again, circling back to this idea of every saturating function looks like an exponential at the beginning, is that the deep learning is currently in a period of rapid expansion, and so we might be coming back to these ideas of interpretability in 10 years or so, and we might be stuck in 10 years ago or so, and the question of how long it'll take us to get general artificial intelligence will seem much more inaccessible. But who knows.

0:21:26.8 VG: Interesting. Yeah, when I think about the whole of human history or something, like 10,000 years ago, things didn't change in lifetime to lifetime. And then here we are today where we have probably been working on AI for under 100 years, like about 70 years or something, and we made a remarkable amount of progress in that time in terms of the scope of human power over their environment, for example. So yeah, there certainly have been several booms and bust of cycles, so I wouldn't be surprised if there is a bust of cycle for deep learning. Though I do expect us to continue

on the AI track just because it's so economically valuable, which especially with all the applications that are coming out.

0:22:04.1 Interviewee: Yeah, you don't have to be getting all the way to AI for there not to be plenty of work to be... General artificial intelligence, for there to be plenty of work to be done. There are hundreds of untapped ways to use, I'm sure, even basic AI that are currently the reason that people are getting paid so well in the field, and there's a lack of people to be working in the field, so there's... I don't know, there are tons of opportunities, and it's gonna be a very long time before people get tired of AI. So yeah, that's not gonna happen anytime soon.

0:22:36.6 VG: True. Alright, I'm gonna switch gears a little bit, and ask a different question. So now, let's say we're in whatever period we are where we have this advanced AI systems. And so we have a CEO AI. And a CEO AI can do multi-step planning and as a model of itself modelling it and here we are, yeah, as soon as that happens. And so I'm like, "Okay, CEO AI, I wish for you to maximize profits for me and try not to run out of money and try not to exploit people and try to avoid side-effects." And obviously we can't do this currently. But I think one of the reasons that this would be challenging now, and in the future, is that we currently aren't very good at taking human values and preferences and goals and turning them into optimizations-- or, turning them into mathematical formulations such that they can be optimized over. And I think this might be even harder in the future-- there's a question, an open question, whether it's harder or not in the future. But I imagine as you have AI that's optimizing over larger and larger state spaces, which encompasses like reality and the continual learners and such, that they might alien ways of... That there's just a very large shared space, and it would be hard to put human values into them in a way such that AI does what we intended to do instead of what we explicitly tell it to do.

0:23:57.9 VG: So what do you think of the argument, "Highly intelligent systems will fail to optimize exactly what their designers intended them to and this is dangerous?"

0:24:07.1 Interviewee: Oh, I completely agree. I think that no matter how good of an optimization system you have, you have to have articulated it well and clearly the actual objective function itself. And to say that we as a collective society or as an individual corporation or something along those lines, could ever come to some kind of clear agreement about what that objective function should be for an AI system is very dubious in my opinion. I think that it's essentially... Such an AI system would have to, in order to be able to do this form of optimization, would essentially have to either be a person, in order to give people what they want, or it would have to be in complete control of people, at which point it's not really a CEO anymore, it's just a tool that's being used by people that are in a system of controlling the system like that. I don't think that that would solve the problem. There are lots of instances of corporate structures and governmental structures that are disenfranchising and abusing people all around the world, and it becomes a question of values and what we think these systems should be doing rather than their effectiveness in actually doing what we think they should be doing. And so, yeah, I basically completely agree with the question in saying that we wouldn't really get that much out of having an AI CEO. Does that...

0:25:50.8 VG: Interesting. Yeah, I think in the vision of this where it's not just completely dystopian, what you maybe have is an AI that is very frequently checking in on human feedback. And that has been trained very well with humans such that it is... So there's a question of how hard it is to get an AI to be aligned with one person. And then there's a question of how hard it is to get an AI to be aligned with a multitude of people, or a conglomerate of people, or how we do

democracy or whatever that's, yeah, complicated. But even with one person, you still might have trouble, is my intuition here? And just trying to have it-- still with the access to human feedback, still have human feedback in a way that it's fast enough that the AI is still doing approximately what you want.

0:26:41.7 Interviewee: Yeah, yeah, I agree. Yeah. I just think that the question of interpretability becomes a very big issue here as well where you really want to know what your system is doing, and you really need to know how it works. And with the way things are currently going we're nowhere near that. And so, if we have a large system that we don't understand how it works and is operating on limited human feedback and is relatively inscrutable, the list of problems that could result from that is very very long. Yeah. [chuckle]

0:27:15.6 VG: Awesome. Yeah, and my next question is about presumably one of those problems. So, say we have our CEO AI, and it's capable of multi-step planning and can do people modelling it, and it is trying to... I've given it its goal, which is to optimize for profit with a bunch of constraints, and it is planning and it's noticing that some of its plans are failing because it gets shut down by people. So as a basic mechanism, we have basically--

0:27:44.4 Interviewee: Because it's what by people?

0:27:46.2 VG: Its plans are getting... Or it is getting shut down by people. So this AI has been put... There's a basic safety constraint in this AI, which is that any big plans it does has to be approved by humans, and the humans have asked for a one-page memo. So this AI is sitting there and it's like, "Okay, cool, I need to write this memo. And obviously, I have a ton of information, and I need to condense it into a page that's human comprehensible." And the AI is like, "Cool, so I noticed that if I include some information in this memo then the human decides to shut me off, and that would make my ultimate plan of trying to get profit less likely to happen, so why don't I leave out some information so that I decrease the likelihood of being shut down and increase the likelihood of achieving the goal that's been programmed into me?" And so, this is a story about an AI that hasn't had self-preservation built into it, but it is arising as an instrumental incentive of it being an agent optimizing towards any goal. So what do you think of the argument, "Highly intelligent systems will have an incentive to behave in ways to ensure that they are not shut off or limited in pursuing their goals, and this is dangerous?"

0:28:53.1 Interviewee: Well, right. It's very dependent on the objective function that you select for the system. I think that a system... It seems, at face value, pretty ridiculous to me that the CEO of a company, the CEO robot, would have its objective function being maximizing profit rather than maximizing individual happiness within the company or within the population on the whole. But even in a circumstance like that, you can imagine very, very, very many pathological circumstances arising. This is the three laws of robotics from Isaac Asimov, right? It's just very simplified objective functions produce pathological consequences when scaled to very large complex systems. And so, in much the same way you can train a neural network to recognize an image which produces the unintended consequence that tiny little perturbations of that image can cause it to radically change its output when you have improperly controlled what the system is doing at a large scale, the number of tiny unintended consequences that you could have essentially explodes many-fold. And yeah, I certainly wouldn't do this. That's certainly not something that I would do, yeah.

0:30:20.6 VG: Yeah. Have you heard of AI safety?

0:30:24.3 Interviewee: AI... Yeah, yeah.

0:30:26.0 VG: Cool. What does that term mean for you?

0:30:27.2 Interviewee: You're talking... What does it mean for me? Well, I guess it's closely related to AI ethics. AI safety would mainly be a set of algorithms, or a set of protocols intended to ensure that a AI system is actually doing what it's supposed to do and that it behaves safely in a variety of circumstances. Is that correct?

0:30:52.2 VG: Well, I don't-- there's not one definition in fact, it seems like it's a sprawling field. And then, have you heard of the term AI alignment?

0:31:00.7 Interviewee: No, I don't know what that is.

0:31:01.5 VG: Cool. This is more long-term focused AI safety. And one of their definitions they use is building models that represent and safely optimize hard-to-specify human values. Alternatively, ensuring that AI behavior aligns with the system designer intentions. Although there are a lot of different definitions of alignment as well. So there's a whole bunch of people who are thinking about long-term risks from AI, so as AI gets more and more powerful. I think the example we just talked about, like the ones where adversarial examples can really change the output of a system very easily, is a little bit different than the argument made here, which is something like: if you have an agent that's optimizing for a goal and it's good enough at planning then it's going to be instrumentally incentivized to acquire resources and power and not be shut down and kind of optimize against you, which is a problem when you have an AI that is similarly as smart as humans. And I think in that circumstance, one of the arguments is that this constitutes an existential risk, like having a system that's smarter than you constituting against you would be quite bad. What do you think of that?

0:32:04.1 Interviewee: Yeah, I was only using the adversarial example to give an example of how easily and frequently this does happen at even the level that we're currently working at. I think it would be much, much, much worse at the level of the general artificial intelligence that would have essentially long-term dynamic interactions with people, rather than a system that's just taking an image and outputting a response. When the consequences of such a system can have long term effects on the health and well-being of people, this kind of thing becomes very different and much more important.

0:32:43.4 VG: Yeah. And like with the problem I was outlining earlier, which is like, how do we get to do exactly what they intended to do? The idea that you have of like trying... Like why would you create a system that wasn't optimizing for all of human values? I was like, wow, ahead of the game there. That is in some sense the goal. So there is a community who's working on AI alignment kind of research, there's money in this community. It's fairly new-- although much more popular, or like, AI safety haw grown a lot more over the years. What would cause you to work on trying to prevent long-term risks from AI systems?

0:33:18.5 Interviewee: What would cause me to do work on it?

0:33:20.6 VG: Yeah.

0:33:29.6 Interviewee: To be honest, I think that it would have to be... I guess I would really have to be convinced that the state of the field in the next few years is tending towards some type of existential risk. I feel like... You don't have to convince me too much, but I personally don't think that the field of study that I'm currently occupying is one that's really contributing to this problem. And so I would become much more concerned if I felt like the work that I was doing was actively contributing to this problem, or if there was huge evidence of the near advent of these types of generally intelligent systems to be terribly worried about.

0:34:28.6 VG: Yeah. That makes sense. Yeah, I don't actually expect computational neuroscience to be largely contributing to this in any way. I feel like the companies that are gonna be doing this are the ones who are aiming for AGI. I do expect them to kind of continue going that way, regardless of what is happening. And I expect the danger to happen not immediately, not in the next couple of years. Certainly people have like different ranges, but like 2060 is like an estimate on some paper I believe that I can send along. It probably won't be a while, won't be for a while.

0:35:00.5 Interviewee: Sure. I don't know, I think that people who understand these algorithms in the way that they work do have in some sense a duty to stand up to these types of problems if they present themselves. And there are many instances of softer forms of AI being used for horrible things currently, which I certainly could be doing more in my daily life to prevent. But for now, I don't know. I guess I just have, I have my own interests and priorities. And so it's kind of a... It's something to get to eventually.

0:35:42.9 VG: Yeah, yeah. For sure. I think these technical AI safety is important. And am I working in technical AI safety? Nope. So like we all do the things that we want to do.

0:35:54.8 Interviewee: Yeah.

0:35:54.9 VG: Great, cool. So that was my last question, my downer of an interview here [chuckle], but how do you think...

0:36:02.3 Interviewee: No, no.

0:36:04.1 VG: But yeah. Okay. So my actual last question is, have you changed your mind in anything during this interview and how was this interview for you?

0:36:08.9 Interviewee: No, it was a good interview. I don't think I've particularly changed my mind about anything. I think that it was good to work through some of these questions and yeah, I had a good time.

0:36:24.2 VG: Amazing. Yeah, why--

0:36:25.3 Interviewee: I typically don't expect it to change my mind too much in interviews, so [chuckle].

0:36:28.8 VG: Absolutely. Yeah, yeah, yeah. Okay. Why do... People tell me they have a good time and I'm like, are you lying? Did you really have... Why is this a good time?

0:36:37.2 Interviewee: No, it's nice to talk about your work. It's nice to talk about long-term impacts that you don't talk about in your daily basis. I don't know. I don't need to be paid to do something like this for instance.

0:36:51.7 VG: All right. Well, thank you so much. Yeah. If you think of any questions for me, I'm here for a bit. I'm also happy to send any resources if you're curious about, like, my takes on things, but yeah, generally just very appreciate this.

0:37:04.4 Interviewee: Yeah, sure. I'm a little curious about what this interview is for. Is it for just you, or is it, like a... You mentioned something about some type of AI alignment group or is there some kind of... I'm just curious about what it's for.

0:37:20.9 VG: Yeah. So I am interested... I'm part of the AI alignment community, per se, although I'm not doing direct work. The people there often work on technical solutions to try to... to the alignment problem, which is just trying to come up with good ways of making sure that AIs in the future will be responsive, do what humans want. And examples of that include trying to build in feedback, human feedback, in a way that is scalable with current systems and works with uninterpretable systems, and interpretability-- certain types of interpretability work. There's teams like DeepMind Safety, OpenAI Safety, different, like, separate alignment community. So I'm like in that space. And I've been doing interviews with AI researchers to see what they think about the safety arguments. And whether... instrumental incentives. And just like, when do you think we'll get AGI, if you think we will. Get a lot of different opinions, a lot of different ways.

[...]

0:38:47.5 Interviewee: Cool. Anyway, that makes a lot of sense and, yeah, I hope that things go well. Thanks for having me. Yeah.

0:38:55.5 VG: Yeah. Thanks so much, really appreciate it. Alright, bye.

0:38:59.1 Interviewee: Bye, see you.