AI and You

Transcript

Guest: Karina Vold

Episode 15

First Aired: Monday, September 28, 2020

Hi, welcome to episode 15! Today we will conclude the interview with Dr. Karina Vold. She is an Assistant Professor at the University of Toronto's Institute for the History and Philosophy of Science and Technology, a Faculty Associate at the Centre for Ethics, and a Faculty Affiliate at the Schwartz Reisman Institute for Technology and Society. She specializes in Philosophy of Cognitive Science and Philosophy of Artificial Intelligence (AI). She is interested in situated views of cognition, the risks, capacities and limitations of current and future AI/machine learning (ML), and the ethical development and use of emerging cognitive technologies. She received her BA with High Distinction from the University of Toronto and her PhD in Philosophy from McGill University. Before joining the University of Toronto, she worked as a Postdoctoral Researcher at the University of Cambridge's Leverhulme Centre for the Future of Intelligence and Faculty of Philosophy. During that time, she also worked as a researcher at the Alan Turing Institute and as a Canada-UK Fellow for Innovation and Entrepreneurship.

Philosophers are coming to the fore now in connection with AI because as we explore what intelligence is for computers we have to know what it is for people, and that connects to more than IQ level but other qualities that go along with intelligence in people, like consciousness, self-identity, self-preservation instinct, motivation or agency, and now we're in the territory of philosophers going back to Aristotle, only it's got a new urgency. Scenarios that were the province of science fiction stories now appear to have direct application to the goals of projects to develop artificial general intelligence.

In this episode you'll hear a reference to something called the *Chinese Room*. That needs some explanation. The Chinese Room is the name of a thought experiment by the philosopher John Searle, from the University of California at Berkeley. You may know that the Turing Test, what Alan Turing called the Imitation Game, is an empirical test to determine whether or not we should consider that a computer is thinking like a human. If, in a blind test of its behavior we cannot tell that it's not human, then Turing said we should decide that the computer is thinking like one.

Searle does not like that conclusion. So his rebuttal starts like this: Imagine a room. We're only seeing it from the outside, where it looks like a shipping container, completely sealed, except for a hole on one side and a hole on the other. Into one hole we put a piece of paper with a question written in Chinese on it. Some time afterwards, a piece of paper comes out of the other hole with the answer, also written in Chinese on it. Searle's question is, does the room understand Chinese? And then he goes on to say that inside the room is a man and a book. The man does not understand Chinese at all. But the book contains instructions that tell him what to do with the pieces of paper that come through the input hole. "If you see a mark with this pattern, then go to this page of the book. If it's followed by a mark like this, go to this page. Write a mark that looks like this on the output paper." And on and on, for many, many instructions. We know that this is a reasonable scenario because computers can answer questions put to them in Chinese by following instructions. You just need a big enough book and a lot of time, but since it's a philosophical thought experiment, the size of the book and the amount of time are not barriers.

Searle says that since the man does not understand Chinese, the room is not understanding Chinese, and that therefore if a computer gives the appearance of answering a question put to it in Chinese it doesn't understand Chinese either, and the same goes even if the question is in English. Therefore computers can't understand anything and they can't think like humans.

If you think that Searle pulled a fast one along the way there, you're not alone. There are many rebuttals to his argument, the most common one being that the system of the room plus the man plus the book does in fact understand Chinese. Furthermore, by using objects that to us have strong boundaries in the real world – books have hundreds of pages, not billions, and people have neither the patience nor the time to follow a trillion rules in such a book to solve a problem like this – he is being so reductionist that we lose sight of the possibility of emergent behavior such as understanding in a system that complex. Yet a system that complex is exactly what we know the human brain is. Similar arguments are made for claiming that evolution could not produce organisms as complex as the ones on Earth right now, but those are made by people who cannot appreciate just how long a billion years is. To be fair, none of us can really appreciate how long a billion years is, but scientists are at least aware of their own limitations in that regard.

A great deal of academic ink has been spilled pointing out how wrong Searle is, mostly because he has never given up defending his argument. In that respect we should at least be grateful that he's provoked so many people to critical thinking by giving them such a large target to throw rhetorical darts at. You might want to try your hand at that too.

At the beginning of the last episode I described the Trolley Problem as an example of where philosophical thought experiments are now being connected with AI development, and you'll hear us talk about that at the beginning of this episode. I said last time that the Trolley Problem is overblown in all the excited journalism speculating about its application to self-driving cars. But there have been heart-rending cases where people have been put in equivalent situations, notably during World War II when there are stories of the Nazis forcing parents to decide which of their children would live and which would die. That once again only highlights how the way we evaluate an individual's response to the Trolley Problem hinges on how much anguish they go through to get to an answer.

And how should we evaluate an AI's response to the other philosophical litmus tests that delineate the boundaries of the inquiries that tell us who we are as humans? You can find out some of that in the second half of the interview with Karina Vold. Here we go.

In AI, some of the most popular philosophical questions occur around autonomous vehicles. There is an obsession with the trolley problem which was brought up long before autonomous vehicles were a twinkle in anyone's eye. And for the benefit of anyone that is not familiar with that, it's a philosophical conundrum that you are positioned next to a switch - one of those long levers that will change the direction of train tracks - and coming down the tracks, you see a train except they always called it a trolley. It's more interesting, sounds better to call it a trolley. So it is a trolley, it's running on tracks. And past you the track forks and you control the switch that determines which direction the trolley will go. Currently, if you do nothing, it will go down one track where some villain who was never specified has tied five people to the track. And you're imbued in this thought experiment with a certain knowledge that it's going to run over and kill all of them. But if you switch it to the other track, it will go over one person that's been

tied to the tracks, presumably by the same villain. And the question is, do you do that? And when we ask questions like that, because there's an entire book of variations on this problem that exposes things about people, what is it that those questions are trying to elicit, to find out?

Yeah, great. Good description of the trolley problem. It's so easy when you have a picture of it. So I think initially, it was the philosopher Philippa Foot who started to use these to elicit people's moral intuitions about intervention, I believe. So whether or not the sort of side effect of an act you do to intervene on an ongoing situation is as bad as if you were to, for example, push somebody onto the track in order to stop the train, or if you were to just allow it to kill somebody. So there was kind of playing on these different nuances to try to pull out people's intuitions on when intervening on situations was more acceptable or not. And it's also useful to pull apart people's intuitions on consequentialism. So the view that an act is moral or immoral based on the consequences of that act, and there are lots of nuances in there. So sometimes it's dependent on the consequences that you can foresee or the unforeseeable consequences. So it was a kind of thought experiment to pull apart some of those intuitions. That was the initial idea. And then of course, when we started to get close to having things like self-driving cars, people thought, "Hey, here's a real situation where a car might have to decide between whether or not it saves the life of a driver and swerves if there's a pedestrian in front of the car, or if it decides to collide into a pedestrian in order to save the life of a group of children or so forth." You can imagine scenarios unbound, I found.

Indeed. And I think people get unnecessarily worked up about this because the vehicles aren't and never will be programmed in that kind of sense. They go through machine learning that is based upon different scenarios and no one who's programming those is sitting there making those kinds of decisions, putting those kinds of inputs in.

That's right.

I think that there may be someone working at one of those labs that may have written a memo to their manager saying, "Hey, let's set up a scenario to train our machine learning algorithms on the trolley problem", in which case, the manager probably freaked out and fired the person of being too dangerous to work there because of the liability that would be created by doing that. But I'm reminded that we judge people by intentionality, and we judge machines by their outcomes. So in judging the outcome of someone running over a pedestrian as a result of avoiding someone else that they were going to run over, we are trained to look at, "Well what was in their mind? What did they intend to have happen?" even if it was a less than optimal outcome. Whereas if we look at the machine in the same thing situation, we assume that it has no intentionality and so we looked to see did it make the best decision based on all the information that we have about that after the fact? Are we going to see any more development of this kind of philosophy prior to establishing that machines have developed some kind of intentionality?

Yeah, so I do think you're right that in many ways, the conversation around self-driving cars and the trolley problem has - and for many good reasons - kind of moved on or moved beyond that. I

think there continues to be a question around how we can get systems to behave, so looking just at the outcomes, like you suggested, in ways that align with our values. So this is called the value alignment problem. And that's not even assuming that the system has any kind of intentionality, weak or strong, but just can we get it to act because we do have systems that bring about modifications or changes in the world, can those outcomes that are brought about align with the kind of situations that we, as humans want to have? And that's a challenging problem for all sorts of reasons that we can go into, but because of that, I think engineers and philosophers are interested in trying to figure out ways in which we can get systems to start to understand whether it's trained on our moral systems are trained on our value judgments. So whether it's a kind of bottom-up learning of human values, or less popular, but a kind of top-down of putting the values into the system, there is still this rich debate going on about how we get systems to best align with what we think is best.

You've brought up the value alignment problem and we'll get to that but you also used another word that was in the category I referred to earlier, which is understand, and I no longer understand what the word understand means. Which was what one of your fellow philosophers, John Searle at University of California, Berkeley, was getting at with his Chinese room argument, which is constructed to "prove" - you can't see the air quotes here - that machines can't think or understand, that only people can. It was interesting to me to find the Wikipedia page on this and see just how many ways people have proven that wrong. Where do you stand on that? I'll explain what the Chinese room is perhaps before the show so you can assume people already know what it is.

I don't know if I have a really well worked-out view on this. I mean, I am sympathetic to the idea that machines don't understand in the sense that we do, but that's not to say that I think it's impossible for them to do that. And I think for me, that kind of understanding though, is more tied into consciousness and having a feeling of aha or a feeling of getting it and less so about just being able to operate on all the nuances or pragmatics of a word, which I think is what Searle was more on to, the kind of full semantic value of a word.

Well, so the value alignment problem, to return to that, which is a very, very hot topic, I don't know if it was originated by Nick Bostrom, but he certainly propelled it to the forefront of a lot of people's minds with his book, *Superintelligence*, which excited a lot of people, most notably Elon Musk. Now, Nick is a philosopher so he wasn't concerned by what would it take technically to build a superintelligence. He just proceeded from the assumption that we had some, and what sort of problems would we face if that happened. And the value alignment problem stems-- Well, it's greatest significant stems from when we have systems that are powerful enough to start having effects on the world that would cause us problems that otherwise you could point to the value alignment problem of a doorknob that's not designed properly, but it's not terribly consequential. But when we have systems that were say controlling strategic nuclear arsenals or global shipping concerns or national power grids, then we do have to pay more attention to that. So the value alignment problem says that we don't even have to have intentionality on the part of the system, it's still capable of just having a bug that causes a

problem. So where do philosophers contribute and what work have you done in that area recently?

Probably the most recent paper, it's a chapter that's going to be forthcoming, hopefully, it's still under review, but will be hopefully forthcoming in the new Oxford Handbook on Digital Ethics. It's a chapter on AI and existential risk, and it's just looking at some of the challenges that kind of go behind that value alignment problem. So spelling out why that problem is such a problem. And I'll just say a little bit about that. So one thing that often gets pointed to is something like what's called the King Midas problem. So the idea of this little folklore of King Midas who wished that everything he touched would turn to gold. But of course, what he quickly learned is that that's not really what he wanted. He didn't want his breakfast to turn to gold, and he didn't want his wife to turn to gold. And so what it illustrates is that even if we could program a system to do exactly what we wanted it to do, we're actually not very good at articulating what we want. We tend to figure out later that there are important qualifications. So that's part of the challenge in the first place of the value alignment problem. And then another problem that we kind of talked about in that paper has to do with moral disagreement. So there's widespread moral disagreement across cultures, but even across individuals within the same culture. And so when we build a system that we want to align with "our values", it's not really clear whose values those are, and that raises some big philosophical problems about whether or not we can assume some kind of value realism. So whether there's some truth that there really are some correct set of values or whether or not something like value relativism is true. And also a further question, let's assume that value realism is true, there's a further divide between whether value monism or value pluralism is true. So value monists, of course, are concerned with figuring out what the one single value is. So maybe something like happiness is the one value that really all other roads lead to. Or it could be the case that there are all sorts of values that are important individually and that some problems don't really get resolved in a perfect way because values sometimes have to compete or get traded off against each other. And so there's not always going to be ideal outcomes for some situations or they'll have to be in critical trade-offs. And then another kind of concern around the value alignment problem is assuming that we could get over those earlier problems if we could somehow get a system to reflect perfectly our current values. Humans, as a society are constantly going through this kind of moral progress. So even if we look back 10, 20 years, our values have changed as a society and things that we thought were acceptable 10 or 20 years ago just aren't acceptable now, and we all see why that is. So we don't want to build a system that entrenches our values in a way that prevents us from having that kind of progress. So there are all of these kinds of challenges that are wrapped up in what's called the value alignment problem. And I think it's the role for philosophers to kind of pull those apart a little bit.

And it's really pointing out, I think, that we don't know ourselves nearly well enough to be able to raise AI in our own image because we don't know what that image is, let alone what an ideal one would be. To take some of the sublime examples that you were going through and give a little more pedestrian one, there was a study done of a version of the trolley problem. It was just asking people "If you had the decision to make, you're driving your car down the road and there are pedestrians in the way. One of them is a young child, one of them is an elderly

person, and you can only avoid one of them. Which one are you going to run over?" And they found that the answer depended upon which country the person was in and that countries that valued younger people - Western societies - would run over the old person and countries like Japan and China would run over the younger one. So I guess it suggests where you ought to move in retirement to avoid the most number of accidents. Now, you mentioned existential risks there. And you were until recently with, as you said, The Centre for the Future of Intelligence, which is, I believe, an arm of The Center for the Study of Existential Risk.

Yes, that's right. They're sort of sister centers, I guess is the best way to describe them. And that's in part because they were sort of co-located physically so the teams became kind of coextensive. And then also, they had overlapping founders. So Huw Price, my supervisor during my postdoc, the wonderful Huw Price was one of the founders of both centres, and Seán Ó hÉigeartaigh, as well, I believe was tied to both. And Martin Rees, I think is primarily at the Center for the Study of Existential Risk, but certainly has had a bit of a role in both centers. So there's lots of overlap between them.

Right. So when an organization titles itself, "The Centre for the Study of Existential Risk," it is squarely putting in its wheelhouse the job of looking at things that threaten the existence of humanity, one of which is AI others would be things like the reversal of the earth's magnetic poles or impact of an asteroid or other things. But what would you say their prime concern or research areas have been? And how were you involved in those?

So my central role was at The Centre for the Future of Intelligence, and CFI as it's called had a slightly different remit. So we were interested in projects around AI and the one that I was on was called kinds of intelligence, which was looking at both ethical questions around AI as well as questions kind of along the lines of comparative psychology. So looking at non-human minds very broadly and how we can make comparisons across, whether it's machine minds or animal minds or human minds in a way that's fair or substance neutral, I guess, substrate neutral. Whereas CSER as you say, The Center for the Study of Existential Risk, is primarily looking at global catastrophic risk and existential risks, including things like biotechnologies, climate change, AI, I think some questions around nuclear arms concerns and probably a number of other areas that I'm forgetting. And so I had some interaction with CSER and in particular, my interests, of course, were the risks around AI. Don't know if that fully answered the question.

Did CSER delegate to CFI the job of studying existential risk from AI?

The way it kind of worked is that CFI which came later, divided itself into a couple different projects. And one of those was on future AI, sort of futures and responsibilities. And that had a lot of overlap with CSER's concerns about long term risks from AI. And yes, I think in particular, looking at things like governance and even global catastrophic concerns, value alignment problem, risks along those lines.

What are the hot topics that are being developed or researched in those centres at the moment?

Well, of course, right now, there's been a lot of talk and interest in what's happening with the current pandemic. So I think that that's been a kind of shift in direction. People are very interested in how technology is being used to either contact trace or to make sense of what we're going through, as well as to sort of find ways to feed into government responses. And then there's also been a lot of work on, as I mentioned, AI governance. So trying to feed into the Global Partnership on AI that's been launched, I think, based out of the UN, between countries like France and the UK and Canada, and many others, I think, European and non-European. So, yeah, feeding into governments has been a major priority. So trying to really implement change. Trying to think what else is going on there. I know my own research better. I'm sure I'm forgetting important things.

Well, let's talk about your own research. What are you going to be doing in the fall?

Yes. So I'll be starting at the University of Toronto as an assistant professor at what's called The Institute for the History and Philosophy of Science and Technology. And it's kind of a department like any other. I'll also have an appointment at the philosophy department in Toronto, and I'll be launching some new courses. So in the fall, I'll be teaching a course on the history and philosophy of artificial intelligence. And then in the spring semester, I'll be teaching a course on some of the ethical and social implications of extended cognition, as well as a course on the limits of machine intelligence. And then I'll just be continuing my research program. So doing a little bit of work on AI and existential risk, some work on neurotechnologies, another area of interest for Elon Musk, and a few papers on the extended mind as well. So that'll be a nice change for me.

And neurotechnologies, I think through popular science fiction, that we've come to the implicit assumption that the mind is the software and the brain is the hardware? Is that a metaphor that you would agree with?

Yeah, I broadly agree with something like that, and that's been around now for quite a while. So it's one of the guiding frameworks of cognitive science is this computational theory of mind. There are some nuances in there about what we mean exactly by software and hardware and the interactions between them, but yeah, I broadly agree with that.

Now, one of the areas where that hasn't previously been something that we could explore, but we now start to think maybe we can when you bring up Elon Musk and things like Neuralink, is that there's a principle in computer science that software can be run on different platforms. What are the implications of that when we apply it to the brain-mind access?

Right. And in philosophy that is called multiple realizability. So this is a thesis that philosopher Hilary Putnam first argued for around 1960 or so. And the idea was that, as Peter had said, we could have the same mental life essentially running on a different substrate or physical substrate. And that also [lent] itself to the possibility that other organisms that have different brains than humans have, maybe it's something exotic like an octopus, can feel pain just like we feel pain, and not some kind of different pain, not like octopus pain, but just pain proper, just like humans feel pain proper. So that relies on this idea that whatever that hardware is or that physical stuff

that that software is running on can be swapped out or be different. Another fun thought experiment is the silicon brain. So the idea that you might, let's say God forbid, but some part of your brain stops working, couple neurons are giving out and so the doctor thinks, "Aha, we can just swap those out for silicon replicas", and so they start swapping out parts of your brain and little by little more and more of your brain turns into just a big silicon chip. Some philosophers will have these different intuitions about whether or not your mental life will go on preserved. This is assuming that the silicon chip really is playing the exact same functional role as what your neurons were doing. So one possibility is that your mental life will be undisturbed. Another possibility is that you'll maybe notice some functional changes or something mentally will change. The other possibility is that you'll start sort of fading out of consciousness maybe so you'll slowly start to lose your mental life as the brain gets replaced. But if you're a cognitive scientist or someone who believes that multiple realizability is possible, then presumably, your mental life should go unchanged. So that's what a lot of philosophers are open to.

And now we can start to contemplate the possibility that that might be in our future, although no one knows how far, but we now have things like the AI called GPT-3, which just became a thing. And it's [an] outgrowth from GPT-2, which had one and a half billion parameters in its machine learning network and GPT-3 has 175 billion, which is roughly twice the number of neurons in the human brain. So at least at that level of organization, we have to start asking what it's capable of. Now, currently, it's not capable of approximating a human being in a sense that would fool many people, but some of the conversations that people have had with it, parts of it would. So we're starting to break down some barriers there and I can see the day happening when we develop something that people are broadly convinced is conscious. Even though the people that created it are equally convinced that it can't be because they know every line of code, but this has managed to establish some sort of empathetic connection with enough other people that they say, "Well, give it its freedom. A philosophical problem in the making perhaps, where some people argue in an imaginary court that this thing deserves the right of self-determination. And then the people who made it say, "No, it doesn't have any of those capabilities" and now the court has to decide, perhaps based on Chinese room arguments, what the status of this thing is.

Yeah, perhaps. It's a bit like the movie *Her* except now, of course, GPT-3 is so much more sophisticated than what Siri or Alexa sound like. So it's even easier to kind of anthropomorphize and to think that there's something hidden behind the layers of the neural network. I hope that the courts don't have to decide things like that but it's possible. I think it's easy for humans, especially when you don't know as you say - if you're not the person who built it, it's easy to look at something from the outside and to start to attribute qualities to it, anthropomorphic qualities. But once you start to dig a little deeper, I hope people will see that we're pretty far from having things that merit rights or legal rights of any kind, especially when we're in a world where there's just so many animals that have far more sophisticated cognitive capacities and that we just don't give any rights to. So this was part of the project that I worked on at CFI was to kind of consider "Just look how sophisticated some creatures are that we don't really appreciate

everything that they can do but we're so quick to attribute psychological qualities and capacities to machines."

And here's the difference between *sapience* and *sentience*. Sentience is the capability to feel, which is what we judge our rights by, whether something is entitled to rights. And yet we are Homo sapiens, we're named after our ability to think.

Yeah. And so sometimes I think we value the thinking more than the feeling.

You were in a project called the Animal-AI Olympics, I think.

Yes. Well, really it was my colleague, Matt Crosby, who as leading that so I was only sort of tangentially connected.

But what was that?

So the Animal-AI Olympics, Matt and his colleague Benjamin, they created a kind of online, virtual playground in which anyone could submit an AI, could build an agent, and submit them to complete certain tasks within that playground. And the tasks were designed based on long-standing psychological tests that had been used to evaluate cognitive capacities in usually non-linguistic or pre-linguistic animals like children or animals to see whether or not you could attribute things like theory of mind to them or different types of capacities like the belief-- The term is alluding me now but the sort of capacity to know that an object continues - object permanence - so that an object continues to exist when you don't look at it, things like that. And so they build these kinds of challenges within the online environment and then let people compete to see how well their agents could perform.

That's fascinating. I believe that humans start from this point that is so close to ground zero that we come into this world without object permanence. So the baby thinks that when Mommy leaves the room, she has ceased to exist, which explains some of the crying. And yet, not only do they learn that which seems like a rather easy thing for a computer programmer to put in their program, kind of assume that from the get-go, but we develop this incredibly sophisticated model of the world so much more quickly than we've been able to figure out how to do in computer software. And did you see any of these entrants in the competition? What sort of progress did you see being made in that area?

Yeah, so I think that there were some pretty interesting results. I'm not as qualified as Matt is to speak on at all but I think part of what his interest in designing some of these challenges was, was to kind of tap into this bigger question, which is now going on in AI, which has been long-standing in philosophy between "how many things need to be built into a system, to begin with?" So kind of innate capacities or building cognitive building blocks for which one can then use to learn and how much of our knowledge about the world just comes from the world itself? And so this reflects this long-standing debate in philosophy between the empiricists and rationalists that was raging in the early modern period. So people like Descartes and Hume had these big debates about that and you now basically are seeing something very similar happening in current AI, where there are debates around whether or not we need to just feed algorithms more information

and more data so they can just start to learn more and more about the world through that data, or whether or not there needs to be more sophisticated cognitive architecture built into the system so that they can get closer to the one-shot learning that we have, as many humans have.

Do you think that by the time the question does arise, "Is this artificial intelligence over here thinking? Is it conscious? Is it human equivalent?", that the philosophers will have a test ready?

Some philosophers kind of already do have tests ready. But the problem with theories of consciousness is that there's just a huge proliferation of views and widespread disagreement. I think there's some general agreement on what consciousness is that we're trying to get a theory of. So philosophers generally are talking about phenomenal consciousness and by that they mean what it is like or the sort of feeling of undergoing an experience, but how one tests for that, in terms of the behavior of a system or the internal workings of a system, there's just widespread disagreement. And what sort of substrates can instantiate that? So whether that requires a kind of neural substrate or not, there's widespread disagreement about that too, and why that would be. So there's at least some theories like the information integration theory that's been advanced by to Tononi, a philosopher, I think based in Italy, although I could be wrong about that, who has a kind of test about how integrated information is within a system but he as well is still reluctant, so IIT theorists are still reluctant to attribute consciousness at least to current AI systems.

We do a lot of talking on this show about how fast technology is developing and how much it might develop in the next 10 years and that happening at a rate that makes it impossible to predict all of the possible changes. If you think 10 years ahead, what do you think might have or hopefully will have changed in philosophy?

I think as a practice, as a theoretical practice, I don't know that much will change in 10 years. But one sort of rapidly changing thing in philosophy is our engagement with the world. I think you were right to say that there's been a renewed interest in philosophy and I think that's true. I don't think that's just a situational perception I'm having of things. And I think that lends itself to an opportunity for philosophers to be less isolated or only talking to each other and more open to engaging with the public and promoting the relevance of philosophy. And so that can lend itself to all sorts of changes within the academic profession. And in part, it's things like training graduate students not to just think like academics or not to just aim for academic jobs, but to think about ways in which we can situate philosophers in labs or in companies and start influencing scientific and engineering practices where they happen. So I think there's a kind of broadening of what philosophers can do. And I think that's achievable in that sort of 10 year period, just based on the renewed interest

It was kind of unlocking the doors to the clubhouse.

Yeah.

Can you see companies having a chief philosophy officer, a CPO and the directors?

Yeah, I think some - I shouldn't say, I think I've heard that this is happening in some cases, or at least, certainly for ethicists. So we've seen already ethicists have done well about placing

themselves in hospitals, for example, where you might need to have some ethical training or make some critical decisions, but some of that's been happening at bigger companies, too. So like Google has been more interested in hiring philosophers. Facebook, at some point, had an ethics team, although they may have disbanded them at some point. And so some companies at least have been showing increased interest in that and whether or not, you know, there's also a lot of smaller companies that maybe can't have a full-time philosopher on board, but there's at least a role for critical input by philosophers. So yeah, I see some potential there and likewise with governments. I think governments are starting to think more about how they use technology and what kind of judgments they use about how much they should be serving current generations versus future generations, for example. And what are the needs of future generations? And what do we owe future generations? And some of those questions can draw on philosophical insight.

An organization's engagement with philosophers is almost a Rorschach test or a reflection of the company's culture and moral values.

Yes. I do remember-- I wish I had someone to cite for this but I think when Norway first found their offshore oil, and of course, they realized what they stumbled upon, they did include some philosophers in their decision-making process about how that money, how that wealth should be distributed, and to what extent it should be distributed to future generations or with future generations in mind. So it's not unheard of for this to happen. That would be a fairly positive thing, I think.

This has been great. We could talk for so much longer. If people are interested in finding out more about what you do, what would you like them to look for?

Well, you can please go to my website, which is kkvd.com and you'll have information about my current research and papers there. And of course, you're always welcome to send me an email and I'll be happy to chat.

Wonderful. Well, Karina, thanks for coming on the show. So much to think about, and we will undoubtedly be splitting this into two halves so that people have a chance to reflect, to assimilate it at a slower pace. You've given all of us a lot to think about. Thank you.

Thank you so much for having me. Thanks.

There you are. I so hope that you got as much out of that as I did, because I find this utterly riveting and so very important to our future on this planet. This kind of inquiry is exactly what I'm here for and I want to make it accessible and interesting to as many people as possible. If you came out of this with many more questions than you had to begin with then I'm satisfied, because that's a primary goal of this podcast. We're raising issues that don't have clear answers and are so far-reaching that what we need to do with them is find more and more questions to ask, because those questions will propel us in the right direction. Humans have never been as motivated by answers as we have by questions.

We were talking there about the King Midas problem, or what happens if AI gives us exactly what we ask for, only it turns out to not be what we really wanted; or needed. You could think of that as the genie problem, because it's like the Aladdin story where you get three wishes, and you make such a mess with the first two that the last one is wishing for everything to go back the way it was to begin with. My two

little girls like a TV show about genies. I asked the younger one what she would do if she had three wishes; figured that she would try wishing for more wishes, which is against the groundrules of genies, if you remember the story. She said, “I would wish for more genies.” Got to hand it to her on that one, I have not heard that answer before; I think she may have found a loophole.

The genie problem, or the King Midas problem, is more formally known in AI philosophy circles as the Value Alignment problem, as you heard us say. How do we get AI to value the same things that we do? In order to do that, it would seem a requirement that we be able to control the AI. And you might think that isn’t so hard. But you’d be wrong, at least when it comes to the kinds of artificial *superintelligences* we can foresee being developed. In next week’s episode, I’ll be talking with professor Roman Yampolskiy about exactly that, what we prosaically call the Control Problem. Roman has a new paper on the Control Problem that’s the most complete exposition on the topic to date. We’ll be digging into that on the next episode of *AI and You*.

Until then, remember: no matter how much computers learn how to do, it’s how we come together as humans that matters.

http://aiandyou.net