AI and You

Transcript

Guest: Karina Vold

Episode 14

First Aired: Monday, September 21, 2020

Hi, welcome to episode 14, and today's guest is Dr. Karina Vold, and she is a philosopher. She got her PhD in philosophy from McGill University in Montreal, and then moved to England to become a research associate for the Leverhulme Centre for the Future of Intelligence (CFI), which is associated with the Centre for the Study of Existential Risk (CSER) in Cambridge. Existential risk means what it says on the lid: Risk to existence. Existence of what? Us. Humans. While CSER looks at several kinds of existential risk, the CFI is focused on exploring the impact of artificial intelligence.

What, you may say, does philosophy have to do with AI? Or maybe you don't, especially if you've been listening to the show for a while or if you've read my book. As we develop our understanding of the universe more and more, as we expand our technologies, we find more and more value where these sciences and technologies start to intersect. While our popular image of a philosopher is someone who wears a robe and sits on a log to think about ethics and morality, the complexity of AI has exploded to the point where we need to think about ethics as we develop.

The most famous example of this is the Trolley Problem, which was originally framed as a thought experiment to explore how people make moral judgements. A philosopher asks you to imagine that you are looking out towards some rails, on which runs a trolley, like an old street car, like the ones in San Francisco. This would work equally well if it was a train, but they called it a trolley and the name had a lot more sticking power than the Train Problem would have. You can see that the trolley is coming down the track, and ahead of it is a fork. If the trolley keeps going down the track it's on, it will run over five people who have been tied to the track. But you are standing next to a one of those long levers that you can pull, and that will change the switch so that the trolley will go down the other track. Yay, you say, let's pull that lever. Not so fast. Down the other track there is one person tied to the track. And this being a thought experiment – no one actually did this experiment for real, philosophy departments don't have that kind of budget – it comes with the certain knowledge that either five people will be killed if you do nothing, or one person will be killed if you pull the lever. What, these fiendish philosophers ask, clipboards at the ready, will you do?

Okay, so that kind of question is all well and good in a lab, but what does it have to do with AI? Well, now we are creating autonomous vehicles; self-driving cars. Suppose one is going down a street and suddenly a kid runs in front of the car. There's only enough time to swerve in one direction and that will take it over a bus bench where a little old lady is sitting. If you or I are in that situation, God forbid, then we have a split second to react and we will be operating on instinct, so no matter what happens, what we do, we will usually be accorded some forgiveness and sympathy, provided we appear appropriately anguished by the outcome. I think that when we ask people the Trolley Problem question, we're not so much looking for a right answer – if you think there's a right answer, the philosophers have all kinds of diabolical variations on that scenario that will change your perspective – as we are wanting some assurance that coming up with the answer causes them some emotional pain. If someone says, "Heck,

yeah, I'd pull the switch" without any trace of anguish then we generally won't trust them. We wouldn't want them piloting our plane, we wouldn't want them taking care of our kids, we wouldn't want them as our doctor. And this gets to how feeling is more important to us than thinking in any agent with enough self-determination to be a significant part of our world. More about that in another show.

But the presumption with a self-driving car is that it is operating at computer speeds, and that it has the luxury of all the time in the world to decide between these outcomes. Furthermore, someone programmed the car to teach it how to drive. To what extent are they responsible for the ethics of the car? Did they teach it which person to run over? Where did that decision come from?

Personally I don't think this is nearly the big deal that it's been made out to be, it's a sensationalized scenario that's catnip for feature columnists and talk shows, but it's an incredibly rare possibility and will likely be a non-event if and when it happens in real life. But now you see an example of how philosophers are being tapped by Silicon Valley and governments interested in AI to help with understanding not just the Trolley Problem but many more subtle but more likely and more useful applications of ethics in technology.

With that kind of epistemological fat to chew on, it's no wonder that this interview spanned two episodes. Karina is now 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 and machine learning, and the ethical development and use of emerging cognitive technologies.

Now one thing we mentioned in here is "winter," and we're not talking about an actual season – this was recorded during the summer – but a so-called "AI Winter". There have been periods in recent history where artificial intelligence was so unpopular that if you were working on AI you didn't admit it. You called it something else, like machine vision or symbolic logic. We are definitely not in an AI winter at the moment, because everyone is rushing to say everything including toasters has AI on board. But the term "AI winter" refers to those periods when AI was so oversold that people were sick of hearing about it and moved on to other things. There have been periods like that in the eighties and nineties. Ok, here we go with the first half of the interview with Karina Vold.

Hi, so here I am with Dr. Karina Vold. Welcome to the show. You're in the actual *AI and You* studio here.

Thank you so much, Peter. Thanks for having me.

Oh, you're welcome. So, how do you describe yourself in a word? Would it be *philosopher?*

Yes.

So you have a PhD in philosophy. Do philosophers say to themselves, "Hey, this is actually the real PhD because it says Doctor of Philosophy, but we're actually the only ones where it's really true"?

Yeah, we do like to say that. Doctor of Philosophy in philosophy.

How did you get into that field?

Yeah. I started my undergrad not knowing what philosophy was and I did my high school diploma here where we are in Victoria and there were no philosophy courses here. So it was initially taking those first courses at the University of Toronto, where I started to learn what it was and started to think, “Oh, this is kind of how I think and the kind of questions that I find most interesting.” So that propelled me to take more classes in philosophy and then eventually ended up with a specialization as an undergrad in philosophy. And then I just kept going and haven't stopped.

Now, at that point, when you started getting into philosophy, and you thought you were going to make a career out of that, get further education and a doctorate in that, did you foresee the kind of interest that would be forthcoming over the last three years that we've seen from technology?

I absolutely did not foresee that. I'm sure some people had a better sense of where the winds were blowing. But no, when I started doing philosophy of mind and philosophy of artificial intelligence, it was during one of the winters and so I wasn't aware that we would have these great advancements in deep learning that would change where we are today. So I've been very fortunate as well, in that so many new jobs have opened up and there have been so many new opportunities in the area that I work in.

So just for context, when was that winter that you were talking about?

So I did my PhD between 2011 and finishing in 2017. So it was right at the end of that probably where things started to get more interesting.

Right. And 2017 is about when all the AI modern explosion took off.

Yeah.

Although we've seen the technology being improved a lot since about 2007 with deep learning, I think, but 2017 was when the popular topic took off and all of the interest in philosophy. So where has that led you that you weren't anticipating at that point? What have you done that you didn't foresee happening?

So much, although I think it's part of a symptom of doing a PhD in philosophy that you don't always plan too far ahead because much of it is sort of up to the job market so you plan to be flexible and plan to be opportunistic. That was my attitude, at least. So going into the job market, I was fortunate to get this wonderful opportunity to work at the University of Cambridge as they started a new Centre, the Leverhulme Centre for the Future of Intelligence, and then that's funded by the Leverhulme Trust, which has funded them for 10 years. And so it was upon getting hired there that I started to work more in philosophy of artificial intelligence shifting away a little bit from philosophy of mind. Even that was a kind of turn that wasn't initially anticipated so that lent itself to all sorts of great opportunities to work with governments and to work with businesses to some extent, and also just to work with more technical researchers in the sciences and engineering.

So did that put you in front of the public or the limelight to a greater degree than one would have expected of philosophers before then?

Absolutely. Yeah, no one was knocking on my door or inviting me to do wonderful podcasts when I was a PhD student at McGill working away in philosophy. So going to Cambridge, there were all these wonderful opportunities to speak about your research and to share what you've been thinking about and to think about how different technologies are going to be used and what kind of applications they'll have for the ethics and social implications. So that's just so many new opportunities to think about, interesting questions that I did not anticipate.

So those fields of technology, computer science and philosophy did not previously overlap, and now suddenly they've collided in this way with people in Silicon Valley approaching philosophers. And I have this fantasy that the philosophers were sitting there going through problems that they've spent hundreds of years working on, not in any great rush and then these technologists come in and say, "Can you solve this for us by next Tuesday? And by the way, it needs to compile to machine code, so we'll check your answer, but don't worry, we'll pay you a lot for it." Has this led to… Well, first of all, correct my fantasy, but what has this done? How have both fields changed? And not only how have both fields changed as a result of this encounter, but how did it start in the first place that technologists suddenly took this great interest in philosophers?

Yeah, that's a great question. So I will say, there's always been overlap in philosophy with different fields in science. So cognitive science emerged, in many ways, in a kind of intersection of psychology and philosophy, and a critique of some of the current views in psychology at the time back in the '50s and '60s. And philosophy and neuroscience, for example, has always had a lot of overlap and intersection in fields like neuroethics. But you're right, that philosophy and computer science, for example, and philosophy and engineering, there haven't really been traditionally these overlaps. Although, artificial intelligence itself, I think, has sort of arisen with a big philosophical history to it, so some of Turing's famous articles were published in philosophy journals. So, in a sense, I think philosophers have always felt like they've had a bit of a stake in AI and that they feel some ownership over that topic. So in that way, it wasn't completely unforeseen but I'm not sure to what extent computer scientists welcome philosophers, so maybe that would be my pushback on your scenario, that I'm not sure how much computer scientists are coming and knocking on philosophers' doors. At least sometimes my perspective is that it's philosophers who are knocking on computer scientists' doors saying, "Don't forget about these important philosophical questions and be careful on how you use some of these rich psychological terms that you use to attribute to the capacities of AIs because we might not be so confident that we can attribute those things."

Right. I think that perhaps in the popular view, the reason for philosophers needing to be involved in AI at the moment is people anticipating the development of AI as having agency, motivation - we're skirting around the topic of artificial general intelligence here, but that we might need to imbue something that is alive now with ethics before it gets out of control. That's

sort of a terminator scenario, which we try and pause our way around here, but is that the level at which philosophy and AI are interacting, or is there a richer, more nuanced effect?

Yeah, that's a good question. There is some of that. So I think there are definitely philosophers feel that there's an important conversation to be had about whether or not we should be describing artificial systems as agents at all, which is a term that I think engineers are much more comfortable using than philosophers. There's a rich history around what agency is and there are many different views, so I'm oversimplifying, but a lot of philosophers tend to attribute or tend to use agency as a kind of shorthand for what we call *intentional agency*, which assumes some intentionality on the part of the system that I think a lot of philosophers would be skeptical to attribute to some of the current systems that we have. And likewise, with consciousness, we would also be skeptical to attribute consciousness to current systems. So some of these terms play an important role in our moral circles - what systems are or animals we include in our moral circles, which lends itself to ethical questions. But I think another area in which philosophy has been playing an important role in current discussions around AI is how we use those systems and in what situations in society we deploy artificial systems. So, as an example, what we've been seeing in the criminal justice system in the US is the use of predictive algorithms to make an assessment on how likely somebody is to re-offend or a kind of recidivism prediction. And philosophers are playing some role in critiquing the use of those systems and some of the questions about "In what cases is that acceptable? What effect is that having on people? How does that compare to human decisions in terms of how we have access to what's actually underlying them, or how we give explanations about why a decision was reached?" So there are some other roles there too, for philosophers, I think.

I see. So the philosophers are basically putting some brakes on the reckless deployment of technology where it has social impact there. You've mentioned the philosophy of mind, I think, a couple of times. Want to talk about that?

So that's sort of started as my core areas, philosophy of mind and philosophy of cognitive science, which you can think of as kind of scientifically informed philosophy of mind. I'm not sure what you want to hear.

Well, your thesis. What topic?

So in philosophy of mind, there are some big questions that tend to dominate, like, "What is a mind? What sort of things have minds, like you and I? What sort of things don't, like the table in front of us? And what are the characteristics that distinguish those things?" But a more recent question that's emerged is about the location of the mind. So "Where does the mind end and the rest of the world begin?" is the kind of question that philosophers Andy Clark and David Chalmers addressed in what became a landmark paper in philosophy of cognitive science. And that was where they defended and sort of advocated for this extended mind view. And that's a view that maintains that the mind can be more than just the brain. So the mind can sometimes extend into the tools and artifacts that we humans use to solve cognitive problems. So just a rough example here is you can think of the Nobel Prize winning physicist Richard Feynman, who used a pen and paper to do a lot of his calculations in mathematics. He would argue with his

historian at the time that his thinking was really being done on the paper with the tool and not just in his head. So it wasn't that he was having a thought and then recording it on paper, it's that he needed that tool to really complete those thoughts and to really have those kinds of thoughts. So that's the kind of idea is that kind of intersection and melding of the mind and the tool or the mind and the machine. That view is what my thesis was on.

Perhaps more currently, you have an iPhone on the desk in front of you that you've brought with you. How much of your mind is in that?

A good chunk for me. A good chunk of my what we call our standing beliefs for me. So information that I maybe would have stored in my head, not so long ago, like phone numbers, or addresses, I'm now storing in my phone. And as a result, I've freed up all this good brain space to do other things. So it's a lot easier for me if I just keep my phone with me and I don't have to remember those things. And that's probably true for a lot of people listening to this podcast.

And we can delegate a lot of that remembering now to Google and Wikipedia.

Yeah. And one other great thing is that now these phones are sort of much more sophisticated than just a pen and a paper, so it's not just memory as a cognitive capacity that they can allow us to kind of extend or offload, but it's things like navigation and problem solving, decision making. So there's a host of different cognitive capacities that are now being extended by these tools.

Extended and performed in the cloud and so now, arguably, part of our minds belongs to Apple and Facebook.

Yeah. So that's where things get really complicated and kind of philosophically more fun is that it also means that those third parties can have an access to your mind that's a lot more direct than it ever was before. So they don't have to go through the standard safeguards of presenting you with information and having you listen and make a decision on your own. Instead, some of that information is being presented to you in a way that you don't even realize that there's a choice being made there on your behalf. So, yeah, it changes people's access to you. It changes access to that kind of information. And it sort of affects your agency potentially in new ways.

That is a take on privacy that most people probably haven't considered before. When I was reading about the philosophy of mind, I was delighted to encounter these characters there. By way of analogy, in computer science, in cryptography, in particular, we have these characters, Alice, Bob and Carol that are used to illustrate stories. Alice creates a message, puts it in a box locks it with a padlock that only Carol can unlock, gives it to Bob, and all of these intricate stories. And really, you can spin an endless story that will get a cryptographer listening for hours as long as you just pepper it with Alice, Bob and Carol periodically, and it can be complete nonsense otherwise. I was delighted to discover that reading in philosophy of mind you have these characters, Otto and Inga doing the same kind of thing. What are some of the adventures of Otto and Inga?

Yes. So there's definitely a rich history and philosophy of these fun thought experiments that kind of illustrate these ideas. And for Otto and Inga, these were the characters that Clark and

Chalmers came up with in their article on the extended mind. So they asked us to imagine these two people. Inga is meant to be just like you or I. She has a well-working internal memory and so when she thinks about wanting to go somewhere like a museum, she thinks, “Okay, I've been there before”, and she recalls in her mind where it's located and then she decides to head in that direction. So in philosophy, we can give a standard analysis of what's happening there, and that's to say that Inga had a standing belief about where the museum was located. She had a desire to want to go there, and then she accessed that standing belief to retrieve the information and then deployed that in order to achieve her desire. So it's a kind of belief/desire action model. So with Inga, what we're meant to do is imagine a kind of-- Well, it's meant to be an analogous case - with Otto rather. So we can imagine that Otto has a poorer working memory and so he records all his information in a notebook and he carries that notebook with them everywhere he goes. So nowadays, you can think of a smartphone, just like you carry a smartphone everywhere you go for most people. And so for Otto, he records information in that and when he needs it, he accesses the information in that notebook and then heads, let's say to the direction of the museum. What Clark and Chalmers argued is that in the relevant respects, those two cases are entirely analogous. And what they mean by that is that they're functionally equivalent, and so the function or the role that the information in the notebook plays for Otto is meant to be analogous as the information that Inga stores in her brain, what that information plays for her as an agent. That's the kind of idea.

Wow, that's fascinating. We're starting to toss around some terms here that fall into this set of things that I and everyone seems to think that we know what those are, but they get tricky when we start asking whether an artificial intelligence could possess them. Things like intelligence, mind, consciousness, self-awareness, cognition. And I've been frustrated in my exploration of these terms, trying to understand them. I know we’re asking way too much to explain in the next 10 minutes, but are philosophers in general satisfied with definitions of things like that?

I would say in general, no, especially intelligence. I think that intelligence is generally seen as a problematic term, a term that has somewhat of a dark and political history or politicized history. It has a history of something that's been used to oppress in many cases and make assessments about different groups, but it's also a term that's become so entrenched that it's difficult to replace. In terms of mindedness, I think there's a bit more of a sophisticated discussion around or sort of a history of discussion around what philosophers mean by mind. So this goes back to what I mentioned earlier in philosophy of mind, one of the big questions is just “What things have minds and what things don't. And in virtue of what do they have minds and in virtue of what do they not have minds?” And so typically, there are two qualities that are pointed to and that's consciousness and intentionality. There's a big, very lively debate about whether or not you can reduce the mindedness to just one of those two qualities. So whether or not intentionality is enough, so you have a mind if and only if you have what's called genuine or intrinsic intentionality. And many people think that it's harder with consciousness to say that because of course, we all fall asleep at night, and we don't want to say that we lose our minds when we fall asleep. So you come in and out of consciousness all the time, but it seems like your mind exists throughout those stretches of time. So in virtue of what does your mind continue to exist or why?

And one plausible reason is to say, “Well, you have all these intentional states that continue to exist even when you're asleep. So you have beliefs and desires and those have what philosophers call intentionality”, or in other words a directedness at the world as kind of a shorthand. So yeah, those features are what typically get discussed in that conversation. And so that's one reason why when we look at machines, we tend to assess them in those terms.

Would I cause distress to a philosopher by asking if she could prove that she was the same person after she woke up from who she was the night before?

For a philosopher working on personhood, maybe? I think there are some philosophers maybe like myself - well I have to be careful, but I do think it's an interesting question about what makes a person a person and especially what makes the person exist over time given the amount of physical and mental changes we undergo. And in society, there are important things that hinge on that. So if we punish somebody for a crime they did as a young person and they're in jail for 20 years, are they really the same person that we're punishing 20 years later? So there are important things that do hinge on that, but for me, it’s never distressed me that much. I've always been more interested in what makes a mind rather than what makes a person, but that's could just be personal interests. Yeah, but you'll distress some people for sure.

Okay, hands up if that was the first time you thought that maybe the geographical boundaries of your mind weren’t what you’d been assuming up until now. As much as the concept of the extended mind hypothesis might sound like a how-many-angels-could-fit-on-a-pinhead, ivory tower sort of academic exercise, when it brings up the question of whether part of your mind is owned by Facebook or Apple it may at the very least clarify our thinking about who that data belongs to. It may illuminate a debate that we’re struggling to develop.

In this week’s look at the latest headlines in AI, Google plans to launch new AI ethics services before the end of the year. According to Wired, this means they will Initially offer others advice on tasks such as spotting racial bias in computer vision systems, or developing ethical guidelines that govern AI projects. Longer term, they may offer to audit customers’ AI systems for ethical integrity, and charge for ethics advice. You may have noticed that some of the biggest tech companies are now calling for increased regulation of AI – well, any regulation of AI. There’s speculation that their motives are less than pure; that they’ve realized that they, the Facebooks and Googles of the world are big enough to be able to conduct Ai development programs under government oversight that would require voluminous regulatory filings that would be too high a bar for small companies to clear; hence, eliminating some of the competition, because Google and Facebook have no problem staffing a giant compliance department.

So what do you think? Is it a good time for government to start regulating AI? What would that look like? Something to think about before next week’s episode when we conclude the interview with Karina Vold.

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