# AI and You

Transcript

## Guest: John Zerilli

## Episode 79

First Aired: Monday, December 20, 2021

Hello, and this is episode 79! In this week's episode we conclude the interview with John Zerilli, Leverhulme Fellow at the University of Oxford, a Research Associate in the Oxford Institute for Ethics in AI, and an Associate Fellow in the Centre for the Future of Intelligence at the University of Cambridge. We've been talking about his recently published book, *A Citizen's Guide to Artificial Intelligence*, which tells you what you as a concerned, capable, and involved person should know about the major issues in AI today. Last week we talked about what that meant, and discussed privacy in particular; this week we'll be talking about bias, and education, and how cognitive science and philosophy inform our evaluation of AI.

By the way, when John refers to fennel near the beginning, some context is in order. He's referring back to the first half of the interview when we were talking about where he says in his book, what if an AI that's deciding whether or not to grant a loan determines from the data that it's been trained on that people who don't like fennel are bad credit risks. It's a great example of the sort of connection that a model *might* come up with that's not illegal, but it just smells bad. Although not as bad as fennel. I don't like fennel, okay, so I was paying attention. Anyway, some meaty stuff ahead, so let's get back into it.

Tell us more about your current project then.

I am looking at a body of law known as administrative law. Some of your viewers may have heard of it. It's the branch of the law that looks at the legitimacy of actions by executive government. For example you can challenge the decision that a minister of the Crown ,or Secretary of State, took in some manner, and you can challenge it on a number of bases. One of the bases of challenging such a decision might just be the minister or the Secretary of State concerned took irrelevant matters into account. So, perhaps in the UK context the Home Office or the Secretary of State, she might have decided to decline an application for a visa on the basis of something like hair color, to use a toy example. Now that decision can be challenged by reference to principles of administrative law, which basically hold members of the executive to account, a high standard of reasonable and fair decision making. In the U.S. is called due process; in the U.K it tends to be called natural justice or procedural fairness. My project is concerned with examining these principles of administrative law and looking at whether they need to adapt to accommodate machine learning in the way that I just said. Perhaps a default position isn't quite the right one to be bringing to machine learning. But I think there are going to be cases on the other side where the machine learning systems themselves need to adapt to the contours of administrative law. So it's not just a case of one yielding or bending to the other. There needs to be some sort of mutual accommodation that fits both ways and that's what I'm trying to look at and examine and get a better grip on in my current project. Because my background is both law, cognitive science, and philosophy, I'm using the occasion to examine

the real marrow of administrative law. What underlies the principles of administrative law? If those principles are ones that we don't want to compromise, say something like equal dignity. So, due process we normally think of that as being the right to a fair hearing based on evidence. You can't just put someone in prison for no reason, you have to have a trial you have to have witnesses, you have to have evidence. If what all that is based on something primitive like our regard for one another as equal beings equal before the law, Then that suggests that any principle of administrative law that's grounded in that primitive principle cannot yield to machine learning. But other principles of administrative law may not be grounded in the same considerations. Or instead, as with the example of the machine that associates a liking of fennel with creditworthiness, perhaps the principle that says irrational decision making is unacceptable isn't grounded in anything other than a sense that somebody's affairs must be pronounced upon in a way that is scientifically explicable.

Doesn't it also mean it's got to be comprehensible to us? Even if the AI is right, if we can't understand it, we can't use it. Is that what you're saying?

That's just the question. Is that what the underlying principles in the administrative law are saying? That a decision needs to be comprehensible to us. I think that is what the principles are based upon, but it's undeniably conditional on another human being making the decision. When it's no longer a human being making the decision then perhaps the default position is misplaced. So insomuch as a human being decides and pronounces on our affairs then they'd better give reasons and those reasons should hold up to common sense, but it might not be a case of common sense being offended or infringed when a machine decides something and uses what merely appear to be irrational criteria. In the case of a machine, it might be something else going on. It's this struggle between requirements of administrative law and what machine learning presents which is the unresolved issue. It's the main question that needs to be addressed.

Right and then complicating this is that the machine is trained on some data which may have bias in it, implicit, hidden, or directly copied from people that gathered the data and there may have been someone biased against fennel in that or somehow that training data represents an inaccurate picture and so the algorithms generate an inaccurate model.

Yeah, that is largely how bias trickles down into affecting real lives. It's because the bias is already out there. The data that is collected simply instantiates in itself the bias, patterns of discriminatory policing for example, and they feed the algorithms and the algorithms are there to detect patterns, so the patterns they detect are going to be largely discriminatory and prejudicial in certain ways and that's what you get fed back.

You mentioned a background in neuroscience; your previous book was *The Adaptable Mind* and what neuroplasticity in neuroscience tell us about language and cognition. And I found it fascinating about how so many fields that weren't previously intersecting have overlapped with artificial intelligence as being the point of intersection. Mostly in the last few years and in particular neuroscience is one of those. Can you tell me how you personally progressed in going from one field to another, how AI entered the picture for you?

So when I took my PhD it was in cognitive science and philosophy and I was entering the job market thinking that's what I would be doing in my first postdoc. But I noticed there were a number of jobs that specifically required people with a background in either machine learning, cognitive science, artificial intelligence on the one hand, computer science on the one hand. And on the other hand, people who had experience of public policy, or political philosophy, political theory, law, jurisprudence. It struck me that this is not a Venn diagram that many people are going to be able to come up against and satisfactorily meet. But my background was in both of these areas. So I thought, well, I'll put an application in, and I got the postdoc and that's how I've been in this field ever since, 2017, when I submitted my doctorate.

What do you think the people who have come to this through a technical background, computer science, engineering and are now driving the field of artificial intelligence? What are the blind spots you think they have that are informed by cognitive science and philosophy?

It's the same issue that was identified by C.P Snow many decades ago when he talked about two different cultures. I still find people that are largely more, should I call them, numerate as opposed to literate, looking down on people that do humanities. And people that do humanities who are more literate as opposed to numerate to some extent looking down on people that just do hard sciences as people that really don't know anything about culture, don't understand all the subtleties and complexities of life. And I think this continues, so that people who go into, call it a a hard science field or at least anyway any kind of natural science, tend to think that the ethical dimension of life, the moral dimension, is really all the soft stuff. So I get the feeling that some people think, well I don't really need to learn that, that's not something you learn in a course; that's just something you pick up as you go in life, that's easy stuff; that's just stuff you get by coping with whatever life throws at you. And the danger is of getting people that are very bright and very capable, highly intelligent people, that go into these natural sciences who end up being funneled in some branch of computer science, maybe electrical engineering, they end up working on building these models without really thinking about the implications of what they're doing. To them it's a technical challenge. To them it is like a Sudoku puzzle. They don't really get any satisfaction from thinking about the ethics of the Sudoku puzzle; they just want to solve the problem. So what can these people learn from moral philosophy, the humanities, more generally is an appreciation for just how complex the moral universe so much so that very acute and highly perceptive and insightful mathematical minds, when they weren't doing pure mathematics or natural science would turn their minds to matters of ethics and find it puzzling and almost just as much of a Sudoku puzzle. I think where moral philosophy meets the natural sciences it is in areas like decision theory, social choice theory, game theory, where you really are modeling moral and political conundrums in very precise and vigorous ways. So I think that if these very intelligent and very capable STEM-type personalities were made to sit through a rigorous course in moral and political theory and game theory, social choice theory - PPE broadly speaking - I think they would be less likely to dismiss all that soft stuff they could see that actually this is challenging and this can exercise my own logical brain in ways that will excite me, and then they can learn to take the moral philosophy more seriously and think through the repercussions of what they are doing, think through all the scenarios that might come. If this system that I'm now developing

and that I've approached as a Sudoku puzzle were actually let out in the wild how many ways could this thing go wrong.

That's really speaking to the culture of Silicon Valley in the drivers that are in it goes back to the early imperative of move fast and break stuff which was the way to get ahead or to survive in the turn of the century. And they ended up breaking the social compact and now we've got hearings where people are going oops, we broke democracy as a result of moving fast in technology, and as you point out they weren't thinking about that. To shift gears for a bit, we are at this watershed moment in history where it's BC - Before COVID and AD - After Disease. So you have an epilogue, you're on the cusp there about the pandemic where you got to acknowledge that. Now that more time has passed - and in COVID era, 6 months is more time – since this is a political book, what sort of connections have you seen with AI emerge as we get further into this pandemic.

Connections between the times that were in?

Things have progressed. For example we're now seeing vaccine mandates to an extent that I wouldn't have predicted a year ago but now those are widespread, causing problems and obviously with privacy implications and so that, since you have an epilogue about the impact of COVID on the situation or how it's changed your thinking, now that you've had more time and the situation has developed more, what would you add to that?

It's underscored the sense that I have that data are perceived to be the answer to everything. More and more and more information. So in managing the pandemic, in plotting or charting its ebb and flow it's all about crunching as much information from as many places as possible. So, people have seen just how powerful that approach is in real time. They've seen not just modeling what's going on and modeling future scenarios but they've also seen breakthroughs in vaccine development which in no small part relied on data and machine learning. So from one point of view, the pandemic has highlighted what was already obvious which is that we're in the information age and everything is coming down to data and extracting as much data as possible. That has other consequences for example, you might call them, epistemological consequences. Just a side note, once upon a time, you might have found a scientist or a philosopher of science avow very firmly that the goal is to understand a phenomenon, is to theorize about it and to understand its underlying properties. The thought that the way you solve scientific problem is just by collecting as much data as possible and putting it into some device and then seeing what happens, that would have been foreign to them. The goal is not to come up with instances, it's to come up with theories and then see if those theories are corroborated. But actually, in this age of information, one way of proceeding scientifically is precisely to do that, to collect as much data as possible and then let the data speak on something like its own terms.

And sometimes the things that we want to find out are not susceptible to human level thinking, like some mathematical proofs have only emerged from computers generating thousands of pages that were beyond human comprehension. The only way to even trust the result was to write another computer program to check it.

So, there's a sense in which then again continuing on with this epistemological point, there's a sense in which science has become in some areas more like engineering and less like traditional science.

So going back to the conversation about the activist goal of this book, let's talk about education not as an application of AI but as educating people about artificial intelligence. How would you like education to shift to do the work that your book is starting?

That's a big question. How would education in the schools, you mean, be changed? I didn't think I'd ever thought about that.

Great, we get a scoop.

I mean, sometimes you hear people talking about making our students future-proof in the sense that they need to have coding knowledge, they need to have technical, STEM knowledge to be able to manage the jobs of the future. And I wonder whether that's the right way of thinking about it, bringing more coding and less of other subjects and prioritizing coding. I think what is probably a better way of thinking about it, the way we take reading and writing arithmetic as foundational, we should take reading, writing, arithmetic and coding to be foundational. And therefore think of it as a non-negotiable that doesn't crowd out other subjects that students could be learning but that's just right there from the start. So students are accustomed to dealing with technology from as early as possible. And it's just part of the equipment.

We're already on that road; but that's going to turn out more of the coders that will move fast and break society. What should we be teaching them to emphasize safety?

I see. I mean I think teaching ethics in the schools is very important. I know that speaking from the point of view of a country with which I have some familiarity, ethics is not a part of the curriculum in Australia. There are moves to make it part of the curriculum and I think this is highly advisable. A secular curriculum of ethical instruction that introduces students to the problems Aristotle and Plato thought about 2,500 years ago. Getting students to think critically about moral conundrums. Obviously, they come up against these scenarios in their readings of literature so it's already there. Shakespeare's plays are full to the brim with conflicts of value. But I think having instruction in ethical thinking in how to systematically approach an ethical problem is probably very valuable. A little bit of decision theory might not also go astray in higher school use. That's how I would think of making students aware of just how important it is to consider the consequences of their actions more broadly, rather than just thinking narrowly in terms of what's in it for me? So the very bright people – to go back to a point we were making before - the very bright people who first invested in Facebook, were perceptive enough and therefore intelligent enough to see that that model of monetizing information for advertising was going to pay off. They saw that. So they invested in Facebook, and they made a lot of money doing so. But they didn't spare any thought about what might come, what's the moral hazard in what we're doing? Now I don't want to come off as so naive as to think that if only they had some decision theory, they might not have made that same error, but everything helps and getting students to think not just in terms of "how things affect me" but "how they affect my environment and everyone in my environment;" that's what I think is missing. I also wouldn't

think that all of this can be chalked up to the school curriculum; these are much larger questions about cultural values, the extent to which neoliberalism has sunk into the recesses of life everywhere. It kind of made us infatuated with the idea of buying property and investing in property and making money. That's a much broader set of cultural issues.

Yes and one of the challenges with this kind of activism is identifying just how many levess you need to pull, because it's never just one so it's not just education, it's not just government regulation, or market incentives it's those and and many other things as well which is what makes it interesting but so challenging. If you were given time to approach, and speak to an educational body, and they said "What would you like to change in our curriculum to forward your agenda here, help us turn out students that are better inoculated against this?" do you have any ideas?

The message has to be gotten out that your actions affect not just you but other people. Somehow or another that message has to be drummed into our youngest people's heads. So because the temptation will always be to think about how am I to advance. Which is fair enough because as organisms our primary responsibility is to keep our own organism running. So it's natural that we will have more thoughts preoccupied with her own survival. But our survival is intimately linked to the survival of others. That message somehow has to be drenched in the curriculum. It doesn't have to be a specific subject, it's more like an ethos that needs to be instilled.

And perhaps it should be experiential. I mean, I was a product of the British educational system where your performance was measured by how you did on examinations and it was just you; there was no component of, how do you do in a group? How might your actions affect someone else? No, you were specifically graded on how you did only by yourself with no interaction, which would have been cheating. And now my daughters go to a Montessori school where it's all about teamwork, so they grow up with a much more educated experience of working in teams. Which I wish I had had because of course you don't do anything in engineering without working on a team and the first time you get on one you are called upon to exercise all kinds of skills that I didn't have in my school either explicitly taught or really explored highly in experiential way, not unless you're counting team sports. So perhaps that's the direction to approach this from. Well, thank you; how should people who are interested in this conversation and what you're doing follow your work and what you might be releasing in the future in terms of other books or other things for them to consume?

You can follow me on Twitter, it's just my name. I often tweet AI-related things and cognitive science and philosophy-related things. If that's your fancy, you can follow me on Twitter. In terms of forthcoming work, I'm engaged in discussions with an agent at the moment about another book that would address similar sorts of issues but perhaps taking a specific angle. I haven't yet settled on what that will be. I've noticed that with every passing month there's now a new book on AI and on big tech and on politics and regulation. So it's soon becoming quite crowded. So, I'm not sure whether I will pursuing this next project with my agent but if anybody has ideas for what sort of book they would like to see in this area, I would be happy to hear from you.

John Zerilli, thank you so much for coming on *AI and You*.

You're welcome. Thank you for having me.

And that's the end of the interview. I liked how we got into how the educational system should change to teach what's most relevant and important in AI. If there are educators listening to this who want to comment or question that, have at it please. I focus a lot on education because I think we're at the stage where the most useful thing we can do to secure a promising future with AI is educate people, starting at a really young age, and I love talking with schools, and right on after that, in the technology and issues of AI, especially ethics. You can find a link to John's book, *A Citizen's Guide to Artificial Intelligence*, in the show notes and transcript.

In today's news ripped from the headlines about AI, Microsoft and Nvidia have built a massive language processor, with the irresistible name of the Megatron-Turing Natural Language Generation model. I mean, how can I not tell you about something called the Megatron when it's actually serious work? It's a transformer – we've talked about those on the show before – with 105 layers and 530 billion parameters. That's three times what GPT-3 is packing. Unfortunately, it's not going to be available for commercial use any time soon. It was trained on a dataset over 800 GB in size known as The Pile – love that title - gathered from Wikipedia and other Internet sources. The downside in getting your education from the Internet is that a lot of the data isn't exactly clean; I mean that in several senses. The researchers said in their blog post that "Our observations are that the model picks up stereotypes and biases from the data on which it is trained." And Microsoft would know about that particularly well because of their experiment with the Tay chatbot which became foul-mouthed and racist after 16 hours on the internet. That's what you get when you train it on Twitter.

I guess my big question about these models is, that we're building bigger and bigger ones that suck down more and more electricity, but we seem to be getting diminishing returns on the intelligence. Like, doubling the number of parameters doesn't seem to double the size of the context they can work with. I could be wrong. Megatron is being used in natural language understanding and, "Demonstrates unmatched accuracy in tasks such as commonsense reasoning." When you hear about AIs and common sense, you do have to realize that it's a narrow interpretation of common sense that's being applied, and so far, it means whatever common sense can be gleaned or mimicked by pattern compilation on the input text corpus. So, to make up an example, if someone somewhere has said that putting out an oil fire with water is a bad idea, a model like this could probably give the right answer if you asked it whether you could put out a grease fire with milk. But only because it had seen close enough answers before. On the other hand, that's how a lot of humans exercise common sense, right? It's just that we can do it on a much larger scale.

We've crossed 40,000 downloads now; in fact it will probably be 50,000 by the time this episode airs. I'm reasonably sure at this point they're not all my mother. Thank you, and if you think about other people you know, I'll bet some of them would like this show too, but they probably don't know about it, so do the show a favor and tell them about it, give us a five-star rating, and more people will come to the party and we'll be able to do more cool stuff.

Like… what's coming next week. We will have a special, end-of-year episode all about predictions for AI for 2022. And you won't have to listen to just me, but this time I'll be joined by a panel of experts in many fields, some of whom you've heard before on the show, and some you haven't.

This might be a good time to revisit my predictions for *2021* as I put in the December 28 2020 show. Let's see how I did:

1. No self-driving vehicle will be certified at SAE autonomy level 5 for use on public roads in 2021. Check, the best we have is level 4 in a very few places. I also said that disillusionment with lack of self-driving car progress would become widespread; but it has not, although you can see that Tesla owners are starting to get more impatient. So I give myself half a point on that one.

2. There *will* be some deployment of autonomous vehicles in narrow applications at lower autonomy levels. Yes, we're seeing companies like Waymo and AutoX deploying level 4 taxis in places like Phoenix and Shenzen. However, I also said that we would see platoons of self-driving trucks on the interstates, and we haven't, so I still only get a half a point. But given the supply chain crisis and the acute shortage of truck drivers, the pressure on autonomous truck development companies to deliver must be *enormous* right now, so… maybe not long?

3. Deployment of narrow AI will explode. Unquestionably this has happened, if only based on my inbox exploding with news of such applications. I don't know if anyone's quantified that. One point.

4. We'll see breakthroughs in medicine due to AI. While we've seen some successes from AlphaFold, I don't think this has been nearly what I had expected, let alone hoped for, so I get at most half a point.

5. Online collaboration tools will develop considerably. Again, not nearly as much as they should have; I speak as a remote worker, and that's really annoying given that we've had an entire year of people continuing to work remotely for the most part. Maybe Facebook announcing the Metaverse as a place to experience through VR headsets would count if they were actually targeting workers instead of gamers. I'm not giving myself anything for that. But if you know of some news that you think qualifies, send it in.

6. Increased deployment of AI in education. Again, I'm very aware of considerable potential and prototyping of AI in education in everything from chatbots to improve recruitment to knowledge bases pinpointing gaps in institutional research, but I've not seen the widespread adoption that there really should have been this year, considering how much universities with physical buildings have been hurting. No points. That was wishful thinking speaking.

7. We'll see major deployments in robotic process automation and customer service. Yes, robot adoption has taken off, and so has the use of AI in customer service chatbots. Not always done well, mind you, but it's there. One point.

8. I said we'll hear artificial general intelligence talked about as a serious research technology rather than blue sky. In some circles, this is true. Bear in mind the prediction is talking about communication of AGI rather than its development. Deepmind unashamedly bills themselves as an AGI company, even though they don't have that yet. In fact, I've heard other companies start to use AGI in their marketing speak, and while I think there's a risk of hyping it to a point where some disillusionment sets in, we're not there yet. One point.

9. We'll see a conflict where AI is part of cyberwarfare. Computer security intrusions have escalated and ransomware has matured its business model – I use the term advisedly – a lot over the last year, so while we haven't had a conflict in the traditional sense, I'll still say that's worth half a point. And finally,

10. We'll see more relabeling of AI applications as something else. I was anticipating that as an extension of the adage that "Once we learn how to do something, we stop calling it AI." I was dead wrong on this; everyone still wants to attach the AI label to as much as possible and none of the bloom is off that rose.

That is a total of five out of ten. We've got to do better this year! So tune in to next week's panel and find out what our collective wisdom says for 2022. That's next week, on *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**