AI and You

Transcript

Guest: Eleanor Drage

Episode 202

First Aired: Monday, April 29, 2024

Hello, and welcome to episode 202! My guest today is Eleanor Drage, who is a Senior Research Fellow at The Leverhulme Centre for the Future of Intelligence at the University of Cambridge and was named in the Top 100 Brilliant Women in AI Ethics of 2022. She is the co-host of the Good Robot Podcast, "Where technology meets feminism," which is very enjoyable and has lots of light-hearted moments with her co-host Kerry McInerney, particular in their episodes where it's just the two of them which they call Hot Takes. Eleanor is also co-editor of a recent book also called *The Good Robot: Why Technology Needs Feminism*. We're going to be talking about all that, plus some quantum mechanics, saunas, ham, lesbian bacteria, and… well it'll all make more sense when you listen. Here we go.

Hello, Eleanor Drage. It's a pleasure to have you on the AI and You podcast. And you also are a podcaster with a book of the same title of your podcast, *The Good Robot*, just out. And maybe you could start by telling us what the book and the podcast are all about and what you aim to achieve with those.

Thanks very much for having me. So it's a great pleasure. And yes, who doesn't love a podcast? We started, as I think a lot of people do, during COVID. And what we wanted to do was to talk to our great heroes in technology like you do. And at a time when everyone was in the kitchen eating cornflakes in dungarees, and you had these great philosophers or to the top people at Google who had just been fired from Google, all at home wondering what the hell they should do with their 9 to 5. And we had a great success from it, just by asking three very simple questions. What is good technology? Is it even possible? And how can feminism help us fix it? And we defined feminism very expansively, so much so that we interviewed Buddhist ethicists or people who didn't necessarily do something specific to feminism. And we just said, "Oh, I think that counts as feminism," you know, we defined feminism very expansively. And very intergenerationally through thinkers that were aged between 16 and 77, who had very different ideas of what technology should be used for and what feminism is, and how those two things relate to each other. And then from that, the book asks those thinkers the same questions, what is good technology, and they respond to those prompts. And I wanted these thinkers that I love so much, and that are all on the bookshelf that lines my rooms, that to have those thinkers be explained in a very simple way. I think that all good ideas can be explained in ways that everyone can understand. And I tried to edit the book within an inch of its life to make it as friendly as possible, so that my mum can read it and have it on her coffee table and ask herself and her friends what is good technology to us? And how can that question help us live our lives with technology in a way that's better and safer?

And here you are, a researcher at the Leverhulme Centre for the Future of Intelligence in Cambridge, and that all sounds like you're engaged in very academic work and must be publishing papers with lots of syllables and long words. And is that true? To what extent does the academic world and the podcasting world, to what extent do they overlap or even conflict?

It’s a great question. And I think that it really helps you understand big ideas if you explain them in many different ways. There’s no one way of explaining something. And this is one of the key elements of explainable AI. You know, the idea that a really complicated technology needs to be explainable to many different kinds of people, its users, the engineers, the people deploying it, people on the ground, affected individuals, as the European Union calls them, that’s you or me. And I think that thinking about all these explanations at the same time can be a bit of a headache, but can be a really useful way of understanding holistically how a system works. So if I’m explaining how an AI tool doesn’t actually observe the world, but creates it, but creates it, I can do this through feminist work based in a gender studies idea that bodies don’t enter the world already made, they are created by the world. And that’s what happens with gender, is that you are born not necessarily as a fully sexed being, but you become gendered through a contact with social norms, through how you’re treated by doctors and nurses. And this is something that intersex people know very well, that are raised, perhaps one gender, and then realize that actually they have the chromosomes of another, or a doctor has made the decision to cut off their genitalia at birth so that they can fit into a gender category. So you’re not born into a gender, you’re made one, it’s about an interaction between a body and an institution. And you can explain the same thing through quantum mechanics. So this idea that AI doesn’t just help you observe the world, but actually actively creates it, can be proved through an experiment that Niels Bohr, a very famous quantum mechanic and engineer thought about a long time ago, which was that if you have these experiments with waves - particles that you shoot through an apparatus - the apparatus and the way that you observe the experiment actually affects the end result. So the way that you observe the world actually affects the materiality of that world. And this is a kind of phenomenal sort of time travel, almost, where the future is decided by the present, is a kind of normal thing. But the fact that the present is just an amazing, mind-boggling thing. But actually, when I ask physicists at Cambridge, what is the most incredible innovation in science, that’s often the one that they tell me. And so you can explain why AI is producing the world rather than just observing it in many different ways. And I really enjoy finding different explanations that speak to people in different ways.

Wow, we got very philosophical very fast here. And I really like that explanation because it resonates with a lot of my philosophy and the reasons why I’m doing this show. And I hope it’s proven through the way that we do it. But I haven’t explained it and languaged it in the way that you just did, which was quite eloquent and philosophical and betrays perhaps some academic background in philosophy. Can you tell us a little about how you came to where you are now?

Through a series of happy mistakes, almost certainly. So I did my undergrad in French and English literature, not because I thought it’d be useful, just because I really like French and reading. And my parents were really pushy they really wanted me to do a conversion to do law. And I wasn’t really sure about it, because I have no attention to detail. And I think I’d have been a terrible lawyer. I think they just thought I'd be argumentative. And that was what it would take, which was mean I'm not sure actually whether I whether I do love to be in an argument, but that's not the story. And I actually bailed on law school the day before, and went to Australia, had a kind of midlife crisis aged 23, and came back to the UK and worked for a tech incubator, because that was a kind of cool thing to do didn't know what you wanted to do, you go work for a startup. And I think to some extent, that's actually still true. And I ended up selling ham online with some computer scientists that I met at a company called TransferWise - now Wise - that do peer to peer money transfers. And so I began in this kind of entrepreneurial way, and ended up doing a PhD in science fiction written by women writers, read through contemporary queer and anti-racist philosophy. And there's a whole chapter now in a book that I published called The Planetary Humanism of Women in Science Fiction: An Experience of the Impossible. There's a whole chapter in there on lesbian bacteria procreation on a Martian colonoid. So if you want to read something really out there, yeah, that's that. And then from that, the sense that I worked for at Cambridge is really interested in the way that AI is narrativized in popular fiction, and in the media. And this actually relates quite strongly to science fiction, because the books that Elon Musk reads, become or influence the technologies that he invests in and creates. So a lot of the kind of fantasies he had in boyhood, become made into real things. Not because the science fiction writers he was reading were especially forward thinking, but because he loved them, and he loved those ideas, and he made them a reality through force of will and investment. So it matters the way that we imagine AI and we imagine the future, and which ideas we're using to bring the future to life.

Wow, I've got a lot of notes here. You said you were selling ham online?

Yeah, I, when I was working at TransferWise, I was doing digital marketing, and I think I was terrible at it. Just bad at answering the phone. There was a company sauna that I spent too much time in, almost certainly. And yeah, so I met these three computer scientists who are working for the company. And they were from the south of Spain, which is quite famous for Jamón Ibérico, this this really delicious kind of ham made from pigs that wander these oak forests and create a delicious sort of marbly textured dried ham. And we began selling that online, which was incredibly difficult because selling any perishable item in a short space of time and storing it is a pain. But yeah, it gave me a kind of the desire to work not at a company to do something where I didn't have someone watching over my shoulder the whole time. And academia is very much that.

So speaking of academia, can you describe your role and work and research at CFI?

Yeah, absolutely. So I'm a senior researcher at this centre that's not a department. And Cambridge is confusing enough because it's split into many colleges and then also departments. And a department is basically a subject that you would study at undergrad. And because no one studies ethics yet, at undergrad, our centre is more a centre for researchers who look into ethical issues in relation to AI, whether they be social issues to do with inequality, or whether they be to do with the way that AI learns in relation to how children learn, or how animals learn. And then downstairs from us, we have the Centre for [the Study of] Existential Risk, who think more about low probability, high impact events, but not just to do with AI, also by warfare, volcanoes and all sorts. And we have two AI ethics masters. One is for people who work full time in industry. And the other one is for postgrad, so people who've just done their undergraduate degree and want a Master's before going into the workforce. So come visit us if you're in Cambridge. And yeah, and we're very friendly bunch.

I attest to that. And I have visited. And I've listened to several of the podcast episodes. They are fascinating. I love the Hot Takes. It sounded like you said you're recording with two of you in a single-person soundproof pod. Is that why they're called hot takes?

Yeah, because it's really warm. Thank you. That is a new angle on that title. Absolutely. We have a fan in that pod that's just really loud. And there is us talking about technology. And we have this absurdly underperforming fan in a very small pod. No; Hot Takes because they're quite spicy. I think my colleague and I have a lot of opinions. She's so sweet from New Zealand, really polite, lovely person. But there's a lot going on in her mind that she kind of keeps to herself at work. And then when we're allowed to talk about tech polemics in the podcast, they all come out, which is kind of fun and juicy.

What are you hoping to change through your work, through your podcast and your book?

I hope to very simply make people more patient with each other. It's very difficult to know how to approach tech issues. And I think if you look at like the polemic around Gemini; so Google Gemini, one of those AI video and image generation tools, and it was critiqued for overrepresenting women and people of colour in its data sets, so that if you typed in "King of England," they might be black. And people from all sides of the political spectrum shouted about this in the pub and on Twitter. And it was an interesting moment because I think as an academic, it's quite nice to not have to solve the problem. That's up to Google. It's not, it's not up to us. But all the shouting was quite upsetting for the engineers themselves who hadn't had any ethics training. They had no idea what to do. So I had Google Gemini engineers say how should I be responding to people who are upset? How should we change the tool? What can we do differently? Which is really interesting in itself, because these, these people had gone into engineering, not into politics, they really didn't have to expect to have to resolve these issues. And there's lots of ways to think about what these tools should be doing. And I think that all the people with opinions need to have like really interesting, productive conversations with each other. And that's kind of what the book is about. It's got lots of different perspectives on what good technology is from very different angles, very different approaches, but they're all working towards the same goal. And by putting them together into a curated book, what we're trying to do is set these different perspectives and conversation with each other. And there's many ways that this can be done, like in a tech union, for example, Apple together and the Google Alphabet are fantastic tech unions, but there's loads. And they're always tricky places to be because people have different perspectives on what kinds of technologies these huge companies should be making and how they should go about making them harm-free; or in kind of tech terms, de-biasing them. And it's really nice to have these solidarity movements that are really difficult, but really important to engage with. So, yeah, my aim is to increase solidarity between people with different opinions on what good technology is.

You mentioned Google a few times here, and it’s a useful place to focus some of this conversation because they have been in the crosshairs of a lot of the heated discussion. No engineer goes in creating the tool to deliberately do what you just described. It’s an accident. But that doesn’t absolve the system that produced those tools of blame, because at least in adequate testing, if they had tested it the same way and found those results, presumably they could and should have fixed them before anything was released, or at least foreseen that that was going to happen.

I agree, but they’re not mistakes, this is the interesting thing, nor are they accidents. They’re just a representative of society, of its inequalities, and also its lack of knowledge about what race and gender is, what representation is, what it means for someone to be represented in a data set. You know, diversity is not just a picture of a black king of England, although, to be honest, I don’t really see an issue with it. And that’s why seeing these things as just errors that can be fixed is not enough. It’s about transforming society.

Right. Thank you for that clarification. They are not errors in the sense that the program was designed to do something else and there was a bug in the code; it is really a consequence of the programming, if you extend programming to include us, to include the training that it gets by virtue of being exposed to our world. And that has caused the problems with representation. And then earlier on, you also mentioned Google when you said you got to talk to some people who’ve been fired from Google, which is one of the areas where Google does not cover itself with glory. So they are perhaps a touchstone for many of these issues. What have you learned about the problems with technology producing these results that cause more division as a result of your conversations, interviews and research?

Yeah, it’s a good question. What I realized is that it’s very much to be expected, because unless you’re creating a tool that actively promotes justice, that is reparative, that’s trying to promote the voices or the needs of marginalized people in the society, what it will just do is perpetuate the status quo. And if that status quo is inequality, which it’s bound to be, then it’s to be expected. So when we talked to Alex Hanna, who was at Google and was talking us through their ethics principles, which are these vagaries; they’re not quite a Hippocratic Oath, that would be better. The Do No Harm thing, though, is actually much harder to interpret in AI than it is in medicine, because harm to who? I mean, in medicine, it’s Do no harm to the bodies that you’re working with. With AI, you have to select which bodies you’re talking about, really. And so don’t be biased. So what does that mean? And when you’ve got engineers who interpret bias as a mathematical concept, that bias is the intercept of the x and y axis? Well, what does that mean to them? Kind of nothing. So I think we talk a lot about, I’ve been talking a lot about Google, but I’m more interested in small- to medium-sized businesses that take this really seriously, they have their own kind of their own culture, their own politics and organization. And they actually want to go a step further than just trying to be neutral or objective, which they know is impossible, and want to work with us to make that happen, which is why we’ve created the world’s first free tool that helps companies meet the EU’s obligations on AI. So they have this big act called EU AI Act, and we interpret it through feminist and anti-racist principles. And that means that you can kind of self-audit, you can be guided through the process of building an AI tool that's going to be the best it can be.

And are you referring here to the tool that's being built by Ammagamma?

Yeah, so it's Cambridge, a team that I lead, and Ammagamma is this really delightful AI company based in Modena in Italy, in the Emilia Romagna region, which was quite famous for resisting fascism, was one of the only regions to stand up against Mussolini. And so they're kind of full of interesting kind of politically active people. But yeah, lovely company, and their headline is, there's no such thing as AI. So they're already thinking quite radically about what AI is. And they're doing the tech side, so they're building the infrastructure for us.

And what do you hope for that tool to be, to accomplish? And what would it look like to use it?

I hope it'll be widely used, of course. And it's going to be, it'll be a kind of step-by-step tool that product managers really are supposed to be using. And so it'll be pitched at the product manager role, the people who have that kind of oversight of the way that AI is being developed in an organization. And it takes you through maintenance data, the ideation stage at the beginning, where you think, is AI even the right task for this? And that's why we wanted to pair with Ammagamma, because they create AI tools on behalf of other companies. And they're very careful, actually, not to create stuff in cases where they don't think AI is the right tool for the task. So they do kind of unsexy, but really important things, like applying AI to make bin collection in Bologna more efficient. And so we create also videos and information about what to do and how to think through that early stage of creating a tool and deciding whether it's the right tool for the job. And then the participatory design stages, plural, where you get in people who'll be affected by the tool, the people who might use it, and you discuss with them what's the best way of creating this? How can we do it in a way that actually responds to social needs, rather than decides them on behalf of the people that are going to use it?

So is this an example of technology being used to improve the cause of underrepresentation?

In what way?

Is the technology helping people make the world better for underrepresented demographics?

I mean, it depends what the tool is, isn't it? Because, again, if it's something that's used in bin collection, I think there's a different conversation there about how that relates to, like, diversity. But still, it's like, where are the bins being collected? Well, where are they not? And which areas? And so the idea of the tool is not just, it's to get people thinking about diversity more broadly than, like, how it relates to things that we might associate more closely with AI gone wrong, like creating an image of a sexualized Asian person. That's something that we imagine to be an issue with AI. But actually, diversity means more than just protected characteristics. It means where the AI tools are working well, and not so well, who they're serving and not serving, who they're putting at risk and not at risk. And we actually do a lot of work in the tool to explain to engineers the relationship between the diversity of the people on the team. So how diversity is not just having tokenizing people by saying, okay, we need one black person, we need three women, but saying, okay, we're building an Internet of Things device. We know these are more likely to be used for domestic abuse. Why don't we consult with or have people on the team that have experience of that? And then we're more likely to create a tool that is going to avoid those situations. So diversity means more than just having people on the team that come from marginalized communities, although that's important, too.

That's the end of the first half of the interview. Next half, next week. I think this is reinforcing for me how our relationship with AI is a kind of "Do as I say, not as I do," because when we train it on past history of human decisions it learns whatever biases were baked into those decisions, along with whatever else we wanted it to get out of that. So when you want it to meet aspirational standards instead of historical ones you've got your work cut out in creating the data that the AI needs, because as we know, it's big data that makes machine learning go.

It's also making me think about Sabine Hossenfelder's video (there's a link to that in the transcript) on why she left academia and how badly she was treated there, primarily for being a woman. Something that's on my mind because I saw this recently. She makes YouTube videos explaining physics problems that are quite engaging.

We have, by the way, had other people from CFI on the podcast, such as Karina Vold in episodes 14 and 15 and John Zerilli in episodes 78 and 79.

In today's news ripped from the headlines about AI, in a decision we should all applaud, Air Canada was held responsible for errors made by its chatbot. Jake Moffatt, from British Columbia, paid full fare for an Air Canada flight to Toronto for his grandmother's funeral after the website's chatbot incorrectly said he could get a retroactive discount for a bereavement fare. That a large language model might make that sort of mistake isn't too surprising. What's reprehensible is that Air Canada responded by saying "the chatbot had provided 'misleading words' " and refused a refund, even though Moffatt included screenshots of the chat. The airline argued it could not be held liable for information provided by one of its agents, servants or representatives, including a chatbot, according to a member of BC's Civil Resolution Tribunal, without saying why it believed that. The tribunal member went on to say that the chatbot is "Still just a part of Air Canada's website. It should be obvious to Air Canada that it is responsible for all the information on its website." Yes, of course it's obvious. Moffatt received damages, and let that be a lesson to every other business that replaces customer service agents with chatbots.

Next week, I'll conclude the interview with Eleanor Drage, when we'll talk about unconscious bias, hiring standards, stochastic parrots, science fiction, and the early participation of women in computing. 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

Get the book: **http://humancusp.com/book2**