Accelerating AI Ethics

Alondra Nelson and Dr Caroline Green

Thick Alignment and the Future of AI Governance: A Conversation with Professor Alondra Nelson

Transcript

February 2026

www. afp.oxford-aiethics.ox.ac.uk

00:00:00 Caroline
Hello and welcome to Accelerating AI Ethics, the podcast of the University of Oxford's Institute for Ethics and AI. I'm Dr. Caroline Green, and in each episode, we explore bold ideas, innovative research, and creative responses to the ethical challenges posed by artificial intelligence.

Today's guest is Professor Alondra Nelson, one of the world's leading voices on the social dimensions of science, technology, and governance. Professor Nelson is the Harold F. Linder Chair at the Institute for Advanced Study and previously served as Acting Director of the White House Office of Science and Technology Policy, where she led the development of the landmark Blueprint for an AI Bill of Rights. She joins us here at Oxford as a Fellow of the Accelerator Fellowship Programme advancing a concept that she calls thick alignment, a richer, more inclusive vision for how societies and technologies might flourish together. Welcome, Alondra.

00:00:54 Alondra
Thank you, Caroline. Good to be with you.

00:00:56 Caroline
It's so good to have you. And, you know, I mean, this is a beautiful introduction, I think, to who you are, but I really put my head into, you know, all that incredible work that you have done along your career and all the accomplishments and just the way that you have created new visions of equality in societies on so many different levels from a research perspective, from a policy perspective.

So, to me, you are somebody who bridges that gap between really serious, excellent research and then impact in society and policy and so on. And I'd like to just now as we start to learn a bit more about you. What is it that drove your career and what drives you now in your work?

00:01:54 Alondra
I think I've always been interested in science and technology. So, you know, my mother was a cryptographer and a kind of systems analyst working on big computer mainframes. And my father was, or is, both my parents are living, is a kind of, a technician effectively. And I think growing up in that household, I was sort of destined in some ways. I mean, all of my siblings work in some aspect of science or technology or medicine. So that gave shape to it. And I was very much a STEM kid. But when I got to college, I became much more interested in the other questions, why do people do science? Why do we pose these questions and not those questions? Why is there inequality with regards to some scientific advances? If we can use science to have better crop yields, why do we still have poverty and communities that are unfamished? So I had all of these kind of social questions. How do you build a laboratory team? All of these social questions. And So when I got to college, it was clear that I wanted to study social science and not science, so I made that sort of pivot. But at university, I also took a lot of science classes, and I was able to take, because I was not then on the pre-med track, able to take science classes that were much more kind of science and society classes, that were much more interesting and much more bridging. And then it was also the case in university, and I know you probably weren't interested in me going this far back, but I think it's very important that I studied as an undergraduate anthropology at a department that required you to study the four fields. So that meant that my anthropology degree required me to do physical anthropology, linguistic anthropology, archaeology, we went on digs, as well as sociocultural. And for me, that was foundational because it was “You can know and learn the science, but you're always thinking about it in this broader context of culture and meaning and values and communities”. And so I think that anthropological founding, that sort of four fields anthropological founding has been really determinative in the way that my work has proceeded, that I didn't have to choose.

I think as sometimes we're told as young people, as students, between science or thinking about culture and community, and that I had this model, an imperfect model, but of a field that attempted at its best to think about them together. So that was actually very important. And so, and I also, I think in the model of, you know, my parents and particularly my mother, had the model of, a black woman who studied science at the highest levels. She was like, she, my mother, was working in cryptography at the highest levels in the 1960s and 70s. And so I also had the great benefit and privilege of not thinking of myself as orthogonal or outside of science and technology. I could, you know, I grew up with my mother driving us around in our station wagon with cathode, cathode ray tubes and punch cards in the back of, it was just like computation was around me, all around me. And so I also didn't, I had the great privilege of never having to sort of cross this Rubicon where I like entered into kind of thinking of myself as someone who could think about or participate in science and technology. So, you know, I think there are things about my trajectory as a child, as a young person, that are distinctive, that make me both really interested and I think more daring than others might be who don't, I don't have a formal degree in science or engineering, and also deeply appreciative of the fact that one can build bridges between the two.

00:05:49 Caroline
Yeah, and what also really strikes me, you know, when I was reading the work that you've been doing and sort of your career trajectory, is that you're also an incredible advocate. for building more equal societies and racial inequality is something that's been a theme, a running theme within your research and your work too, hasn't it?

00:06:12 Alondra
Yeah, for sure. Yeah, I mean, it's, you know, I think if we can imagine and have visions for society that include more of us, that empower more of us to live in the fullness of our humanity, and I think this as a scholar and as just an individual citizen of the world, like what are we even doing here, right? And so those commitments run very deep through all of my work. And I know to some people it feels or reads as striking or out of the ordinary, but for me, like it's, they're all of a piece. I don't know how to think about, you know, I've started a lab at the Institute called the Science, Technology, and Social Values Lab, and that's really, about, what are our aspirations for innovation? And shouldn't those aspirations be, more justice, more inclusion, more equality for more people? Like if we are going to turn the highest powers of our minds and our innovation to things in the world, shouldn't it be those things? Yeah.

00:07:23 Caroline
So what I also find incredibly striking about your work is, that, and you've just also explained that again, how in your life you have been somebody who's been thinking in a very interdisciplinary way and, bridging kind of what might be silos at times. And I think you've done that really successfully in various ways. And, you know, working in AI ethics, working at an institute for ethics in AI, where we're an interdisciplinary group of people, from an academic point of view, but also wanting to engage with people outside academia, this is always a question on how to do that really well. And so I'd love to hear more from you, how you see this interdisciplinarity, this bridging exercise going well. What can we do as researchers to be successful?

00:08:15 Alondra
I would say in recent years, we have lessons, very material, dramatic lessons about what happens when we don't have interdisciplinarity and why we need it. So if we think about the pandemic and we think about all of the consternation and controversy and polarization around things like vaccines that have happened ever since, we don't know, it will be historians a generation or two or three beyond us who are able to give us clarity about exactly what happened over the last few years. But we certainly do know that we had some of the world's leading scientists making some quite revolutionary new, you know, using a new platform to create vaccines and not having an ability to convey to the kind of global community, why it was okay that they were created so quickly, why it was okay that we were able to decode the genome for the virus, for SARS-CoV-9, and very quickly, we were able to have a vaccine in less than a year. I mean, this was an accelerated pathway for how we do science. And people aren't wrong to say, is it safe, it's very fast, et cetera. And I think we really failed globally to have a conversation about what was happening. And I think to the extent that people, scientists, policy makers, people in public health, attempted to have that conversation, it was in their language. It was not in the language of the broader public. So it did not say, I completely understand why you might be concerned about what this vaccine might mean for your family or your child or your grandmother, et cetera.

And let me explain it to you in a way that, you know, makes sense. And that not only tries to, you know, like make you more literate about vaccines, right? But that tries to understand the perspective that would make you cautious about wanting to use them. And I think so much of the communication did not include professional communicators, did not include social scientists, and people who, and actually just regular people who are non-experts, who could come at those questions from people's fear, from their values, as opposed to if you could only understand how a spike protein works, then you would understand why this vaccine is effective. And for me, that was a great opportunity for interdisciplinarity that we missed. That might have put us in a better position in this moment. It might have, you know, might have had some mitigating effects with regards to some of the polarization we're experiencing, certainly in the US, but in other places around sort of public health issues, for example. So I think for me, and I'm going to probably say this a lot over the course of our conversation, but when you focus on the outcome and then you reverse engineer from there, I think interdisciplinarity is often, for me, how you're reverse engineering. So if the outcome is we want a healthy population who trusts that these drugs can work and that appreciates that things have moved more quickly, but that it is a good thing and that we have taken, necessary protocols, then how do you get to that trust and how do you get to that place? And that requires lots of different perspectives at the table about any given thing.

I would say the other piece, particularly at this moment in my career, that makes interdisciplinarity so important is, you know, as I was talking about the bridging between being someone who loved and worked in science and being someone who very much appreciated the social science of it. I think I also understood and was able to, I've tried to do uniquely in my work, I've done work around genetic diseases like sickle cell anemia, and it's really my work on human genetics that becomes my bridge to thinking about big data and then AI. But it was always very clear to me that if I was going to work on sickle cell anemia, I needed to understand the genetics and I needed to understand the change in technology over time. I needed to understand what was going on in scientific fields. So there was kind of a history of ideas. There was a little bit of history of science and medicine. Also what was happening in communities, as well as what was happening in sort of the technical, the sort of biotech of genetic disease in the 70s and the early aughts and the time periods that I was writing. And so the Interdisciplinarity is also about wanting to understand the thing I'm researching in its fullness, and then wanting to be able to pose questions about a research project that are the right questions for the project, that are not just the questions that I bring because of my, I have one discipline and one field and one hammer that I bring to sort of every data set and every context. I think I'm always interested in having a much more organic and inductive relationship to research projects that are like, what are the interesting projects about this domain, which may not be the interesting questions or the questions I think will have the most impact if we answer them or illuminate them in another context.

00:13:38 Caroline
So I want to actually, I think the COVID-19 examples and public trust around vaccinations, that's really interesting because we're now seeing issues around trust and AI in communities too. So I would like to take this conversation further in a minute, but just to end our conversation around interdisciplinarity. So I would say I'm an interdisciplinary researcher too, you know, I've got a law degree in gerontology and so on, so all sorts of a mixed bag. But I guess there's still a lot of value in having deep expertise. But then being able to have meaningful conversations, open conversations, and communicating ideas in a way that other people outside your discipline can understand. Would you say that?

00:14:25 Alondra
I agree with that. I mean, I also think done well and correctly that interdisciplinary work is harder because you're not just, you have a depth of expertise, of course, but what one doesn't want, I think, is just a bunch of dabbling in a lot of other fields, right? And so in your case, it means that you have

degrees in two very, people would think kind of disparate fields. And so what I was trying to get at was knowing the sociology and also knowing the science, right? And having to bring depth to both of those, but also appreciating that you can, and this is where I think collaboration, which is what I would add to this, is really important because you can't dabble in all the other things that you might want to know about a particular research topic, particularly like something like AI. And so then you also need other people who are willing to bridge across disciplines and work in collaboration with you. And I think AI is a perfect example of that.

00:15:23 Caroline
Yeah, I agree. I think we're seeing this now too. So when working with communities, people in the community, people who use technology every day, but might not understand in much detail how it works. They're soaking up news cycles around existential threats and so on with AI. And there is a lot of public distrust. And then there are the experts, the computer scientists, who are working on AI and sort of developing it rapidly. And here I really see a space of what you've just been talking about. So how do you see this as playing out?

00:16:10 Alondra
I think it's a very much a parallel example. I mean, I think if we think about the introduction of ChatGPT, so we'd had GPTs before, but ChatGPTs introduced in November of 2022. And there's a market race dynamic happening. So one can somewhat appreciate why a company would release a technology that powerful all at once to the public without advance notice. But you can also appreciate that it wasn't exactly a great idea. I mean, you know, certainly just in the US context, I mean, what that unleashed was, schools banning ChatGPT, it was just a created a kind of panic and that we're still living in some regards. And so I think when you have entrepreneurs who are involved in, stiff competition with other venture capitalists or other private equity funding groups, or you have scientists who might have lab competition or who are not adept in speaking to the public, I think the distrust that we are living in, particularly in place around AI in places like the US and the UK, are direct results of that.

So we know from the Pew Research Center and things like the Edelman Trust Barometer and other waves of data since 2022, that levels of mistrust, particularly in what we call the West, the US, UK, et cetera, are high and expanding. And part of that is an outcome of releasing a product and then saying, we're all going to be extinct. It's so powerful. We can't stop ourselves. We won't pause. And it's really, really dangerous. And I think the lack of self-awareness on the part of well-meaning scientists, well-meaning computer scientists, machine learning experts, around that kind of narrative that was just thrown into the public sphere without appreciation for the fact of what it might unleash. And it has unleashed exactly that. Like, you know, fear about people's futures, about their jobs, about their healthcare, about the mental health of their children and people that they love. And so I think it's It's, if we want higher levels of trust and if we want for good uses of these tools and systems to be adopted, we really have to be much better at communicating about them and ensuring that the public has a role and a voice and a seat in conversations about how they should be used, where they should be used, et cetera.

00:18:59 Caroline
So I think that leads on really nicely to this incredible piece of work that you've been leading in your role in the White House, the blueprint for an AI Bill of Rights. And so you joined us earlier this year

for a conversation at Westminster Parliament on aligning AI for human flourishing. And you went into some of the methodology that you, know, how this whole blueprint came about and was absolutely fascinating. Could you take me through that a little bit again? How did it come about?

00:19:29 Alondra
So I worked in the Biden-Harris administration for just over 2 years, and we came into office in the middle of a pandemic. And the question for us, I was working in the leadership of the White House Office of Science and Technology Policy, was, how do you conceive of making science policy in a context of low trust? And in the high water mark of a pandemic. Like how do you even begin to think about that? We know how we typically think about that, which is you get all the eggheads and the nerds and you put them in an office in the White House and we do kind of nerdy egghead things. And we are kind of closed off from even the rest of the White House often. And somebody would call up from the West Wing and say, What do we know about X? Kind of often, very kind of arcane technical or scientific thing, and somebody would scurry up with a memo or something and answer those questions. But this was a different moment, and I think a moment like we hadn't had in a very long time.

And so we had been watching the fact that we were doing very poor public communication around the vaccine rollout. We were very aware that there was, there were viral videos about things like facial recognition technology and the ways that they worked and didn't work for different communities. These were going all over social media. So there was a kind of growing, there were the issue, these issues that were a concern. And then there was a whole suite of other issues. I mean, the Office of Science and Technology Policy has this kind of very vast portfolio that goes from like agriculture to zoology to nuclear to national security to education. Like it's kind of, and how science and technology kind of cross-cut all of those. So we knew that we, needed as an administration to say something about AI and to address how the administration was going to approach and what was going to be its one of its first kind of approach, the way that they communicate with the public about its approach to AI. And we also knew in the context that we were in that science and technology policy makers just had to get better and had to commit ourselves to engaging the public in our work, full stop. And we'd had some tools from the Obama administration had started doing challenges, and they'd created other kinds of sort of platforms that allowed the public to engage the Office of Science and Technology Policy and other agencies. But we thought we had to do, we could use some of those, but we needed to do more. So we set out to do a year-long kind of engagement with the American public around AI, their aspirations for it, their fears about it. And that included, it began with an op-ed that we wrote and wired, myself and another OSTP leader.

That was a kind of call to the American public, says, you know, what are you, AI is fast moving. There's chances that it might threaten some of our rights, some of your access to opportunities. There's chances that there might be very good things. What is the American public, how should government should be thinking about these issues and what are your concerns? And that op-ed ended with an e-mail address to the White House. So you could just write to the White House. And somebody was checking it regularly. We got fewer emails than one would have expected. We thought we were just going to get spammed or some other. But it's also the White House, so I'm sure there was quite a significant spam filter. There was probably a lot that we did not get as well. And it announced that we were just going to have this process that was open, that was not just us top-down making decisions about a powerful, transformative technology in American and global society. We also had, as part of this, lots of roundtables about different top topics. So we had, this was still, again, still only just starting to come out of the kind of high COVID moment. And so we, which brought with it the opportunity to have online kind of workshops and online conversations. So we did lots of those with civil society leading the way on different topics. How do people want to think about privacy issues in AI? Should we be thinking about healthcare issues in AI? So we had these kind of topical issues.

We also, this project was led out of my team, had something we called office hours in which everybody who worked as a member of the team had to set aside two hours every Tuesday on their calendar that they had to block out. And that we would talk to anybody who wanted to talk to us. And that included often colleagues from industry, it included often colleagues from academia, but it also included just regular people. I mean, we had meetings with rabbis, we had meetings with high school students who were telling us how they were thinking about AI, using AI, their concerns, their hopes. And we did that for not quite a year before we had a draft document. And What the document really is, kind of distilling kind of best practices from people that we engaged with. I mean, what you don't, I think it's this is probably maybe kind of distinctly American perspective, but we do appreciate that innovation comes out of the academy, but it certainly comes out of companies and company AI labs as well. And so, we were not telling, labs anything that they didn't know or elite academic researchers, but we were trying to learn from them and others sort of what might be some sort of principles for how we think about how we can deepen engagement with AI in American society. And these included, you know, very basic things that you should have some modicum of data privacy, that AI systems should be safe and effective, that you should know if AI is being used for a consequential decision. So if this is a decision that's about a mortgage loan or health care or job, something that just really actually matters for the trajectory of your life, you might want to know that an AI system was used. And if there's a decision that's made that gives an outcome that you don't agree with, There should be, some recourse. You shouldn't just be caught in the version, in an AI version of a telephone tree, which you just kind of press 0 until you get more angry and then you press 0 harder and no one comes on the line. So those were very basic things.

I mean, it wasn't, there was nothing there that anyone hadn't said before in the AI policy space, but it was important that the White House was saying it, like so that the sort of bully pulpit of the White House to say these are going to be some guiding principles for that work. And when we launched it, we invited members of the public, folks from civil society that we had engaged with, in addition to industry and academic leaders in AI who would have been included in any administration. And I think the last thing I would say about the process is that we worked very hard on the communication of the document itself. And the subtitle of the document, so it's the White House Blueprint for an AI Bill of Rights, And the subtitle is Making Automated Systems Work for the American Public. And it is typically White House policy documents are, if not quite navel gazing, are addressed to other policy people or to a world of policymakers and think tanks and civil society. And we wanted this document to be addressed to the public. And so it is free of jargon. It does not assume that you have any expert knowledge of AI. And I think to go back to where our conversation began a little bit, it really starts with the outcome. Like you want safe and effective systems. That's the very basic element, fundamental thing that we should want from these systems. And then says, how do we get there? What do we have to do in policy? What do we have to do technically to arrive at that place? And we say it in just very plainly stated language. And I think that for us, that was very much a lesson from the miscommunication, I think, of the pandemic.

00:27:39 Caroline
An incredibly valuable document in these insights. You were just saying how people, they've got a call they have to do to their local authority or whatever, and then a chatbot comes and they can't get the answer and they keep on pressing. And I mean, it's a very real experience, right, of so many people. Also people I work with or I've spoken to have told me, look, I'm just trying to make an appointment with my GP. and I can't even talk to a person anymore. It means that in worst case scenarios, that people just don't go and see their doctors anymore. So having these experiences so much now at the foundation of this document and these insights, I think that's incredibly valuable.

00:28:22 Alondra
Yeah, I mean, I've been watching, you know, we've been tracking me and some of my former colleagues, how the document's been traveling because the Biden administration is no longer in office and there's a new kind of vision for AI. But the document has continued to be used. It's used in school curriculums. It's used as a Harvard Business School case study. But it also, I think, is, you know, I call it kind of a kind of civic infrastructure. It tells people to have that document and to know that the document exists, that they're not wrong or crazy to say, I'm trying to get to my GP and I shouldn't just be kept in this kind of, Seventh circle of AI hell, and so I think it just leaves open the possibility that there could be another way in which we engage these tools in our world that doesn't look like the one that many of us have, and so I think... regardless of who's in the White House now or in the future, it just created different expectations for our engagement with these tools. And I think that's important.

00:29:25 Caroline
Yeah. And also, I wonder, you know, you say you've been engaging with all of these people. You see that I sometimes feel that there is sort of like an assumption that, you know, people in the communities, people going about their lives, that they're not really interested in AI or technology and they're using it, but they're just sort of like subject to it. But actually they've got a lot to say. They've got a lot to say about it.

And so one of the, when I read through some of the work, your past work, I came past this, a piece of writing from Future Texts, 2002, I think you published that. And I found it really striking. So here you were talking about the concept of digital divides. So that's something that I come across in my work a lot, specifically for older people. So aging, there is this idea that older people are often disconnected from technology, which I would say isn't actually true. So I immediately read into this. And what you say here is that there's sort of like digital rights, that concept means there's an overemphasis on the role of access to technology in reducing inequality as opposed to non-technical factors. So I'd like to pick that up with you, that idea of the digital divide. and the role that it plays now with AI and across communities or groups of people where we feel like that might be an issue.

00:30:57 Alondra
Yes, I mean, it was, that is a phrase that comes, I think it's coined in like the late 1990s. It might even be coined by the NTIA, which is the National Institute for National, something, Information Technology Administration in the US. So I think there was a report issued might have been in the Clinton administration that sort of coins the phrase digital divide. And in that essay from 20 years ago, can you imagine? I was trying to sort of capture the fact that it had become this kind of almost this self-fulfilling prophecy that like you have technology, some people have access, some don't. And

then sort of that's just the way of the world as opposed to imagining that there were ways to have, that we could have ways to have access to technology, ways to have, I mean, to education and other ways in which technology was being used that didn't require that you could only do it through technology. So digital divide was sort of becoming how do I want to say, almost a kind of polar thing that was being created very early on as like sort of the other phrase that's used is sort of the haves and have-nots, as opposed to there being a curiosity on the part of researchers or an interest in. Sometimes people make choices not to use technology. Sometimes people are using some technologies and not others and have very clear reasons and explanations for that being the case.

There's some research, for example, that suggests that in the United States, after Hurricane Katrina, part of, many well-meaning folks went there to really try to help rebuild this community that had lost so many lives and so much infrastructure. And some of these were innovators in the education space who wanted to bring all of these technology to young people. And so there are some, you know, sort of qualitative research about schools in New Orleans in which all of this computation was entered into the school system. And the outcome and the data shows that the young people didn't feel cared about. They thought that they were, instead of feeling like we had sent our best technology to give these young people the best educational experience, what they reported back was no one cared about us to give us teachers. And so we got laptops and iPads. And so I think that that's an example of wanting access to bridge the question of, or the issue of educational equality or educational inclusion, and thinking that if you just add the computer, the thing is resolved. And I think now that is compounded with AI, because we really think if you add the thing, if you add AI, all of these issues are going to be solved. And so I guess it, you know, the digital divide, my kind of critique and engagement of it in that essay, Future Text, was really I think an encouragement for us to just not to think in socio-technical ways and not narrowly technical ways to the extent that we think that technology can be helpful with social problems, and sometimes it's not helpful at all, and we've got to leave that open.

00:34:12 Caroline
Yeah, incredibly important point, thinking also about, again, going back to the area that I research in. So aging and how technology can help here. And where sometimes we're seeing kind of like a tendency of, well, we put some sort of AI or technology into somebody's home and then they can stay there more independently and autonomously and technology will help. And that may be the case for somebody in a specific time, but then another day in another situation that might not actually work or So I think these nuances that here that you describe and thinking about it in more depth, I think is amazing.

00:34:50 Alondra
And it's so great that you're doing this work. I mean, so few people are working on aging and technology and certainly AI and technology. It's really important, particularly as we think about the possible expansion of things like social care robots, right, that are very big in societies like Japan. But as we have graying societies and fewer care workers. This is, I'm just telling you things that you know much better than me, but it's an issue we're really going to have to think hard about, all these nuances.

00:35:19 Caroline
Yeah, I think so, and I can definitely see that there's... more engagement starting to happen with

aging populations and what we'll need here. And I think what you just said with care bots is one of the kind of like panacea to issues in care, which I think there's again so much nuancy and we can lean into some of the work that you have done also to engage with older populations on how technology is working for them and how the narratives need to be stretched here. I'd now like to go into your work that you're doing here specifically with us at the Accelerator Fellowship Program. Could you tell us a little bit about that?

00:36:03 Alondra
Yeah, well, first let me say I'm just, I was so delighted and honored to be invited to be an inaugural Accelerator Fellow. So I'm really so grateful to be a part of this community. And I, you know, I'm doing a couple of things. One is a project called Thick Alignment that I started actually partly here, I mean, not only here, but that I began to elaborate here in a lecture I gave at the Institute for Ethics and AI a couple of years ago, and is trying to think about what's called the alignment problem.

So there's a very important book by Brian Christian about the alignment problem and about AI development and research and the way that researchers in computer science have been trying to think about how you get AI systems to do the things you want them to do? And moreover, not just that, so that's kind of a technical question, solely, how do you get them to do what you want them to do in a way that's aligned with human values, whatever all of that means? And so I was very... interested in that conundrum. And also, I think, going back to where we started, because it is a question that I think in its fullness is about how you bridge the technical with the social or the scientific with the social, which has been the question of my, you know, it's been something I've been kind of interested in across my career. But what was very clear is that how the sort of work of alignment was getting built out significantly in a CS subfield or an AI subfield that's called AI safety was just quite narrow and technical and actually wasn't really about alignment at all in a way. And so I really came to think of it as thin alignment as just the like bare minimum, like not even beginning to get close to how we need to make sure these systems work and how we need to have a more kind of dialogical process with regards to building, developing, deploying these systems if they're to work and work well in society.

And so thick alignment for me borrows from lots of scholars before me, but certainly Clifford Gertz, who was an anthropologist at the Institute for Advanced Study, where I'm on the faculty. He was actually the founding faculty member of the School of Social Science that I'm at the Institute for Advanced Study. And he has a famous essay called, a concept called thick description that for him comes out of an ethnographic approach that is much more inductive. It's sort of what are all the things, you know, describe all the things, try to, before we think that we understand or can explain, like can we just describe, can we just illuminate the sort of relationships, interactions, the things that are happening around us? So I draw on Gertz to sort of think about thick alignment as how do we, if we're going to align a technology, and we're sort of seeing examples of this.

So if we think about the challenge that we're seeing with young people becoming besotted and kind of seduced by chatbots, in some cases leading to things like suicide and, you know, self-harm, like very, very dangerous. Those chatbots have been, to some degree, aligned, right? They have gone through probably some bare minimum, I don't know, process of safety testing and alignment in the context of race dynamics, in which people are always trying to ship product as quickly as possible because you're competing with other companies. But I think for the general question of sort of thin alignment, sure, we think people will put this range of inputs and you'll get a pretty reliable set of outputs. What we're seeing is that, you know, chatbots are being used for myriad other things. People are using them for therapy. They're using them for companionship. And I think a thick alignment approach anticipates that these are very human natural things to do and sort of says, we're not going to be able to solve this, but we're going to at least be able to, I think, pose, take our time and pose much more seriously questions about how we're going to think about this when it happens.

So an example of this, you know, I've been writing quarterly for Science. And I talk about my most of the, a column that I wrote is called the ELSI for AI. And ELSI is an acronym that stands for the ethical, legal, social implications of, and it starts with the Human Genome Project. And so I'm advocating for how we might have one of these for AI, not for an answer, but that it might give us some ways of thinking about how to align technology with anticipating social uses, social harm, social risks that we might need to think about. And it's not, I'm not trying to suggest that if we have a more value-centric, a more dynamic, you know, complex way of thinking about alignment, that we will never have risk or harm, but that we at least have, we've at least thought about it. So in the ELSI case, I think, one successful outcome was when you had the case of a Chinese researcher who did human genome editing, gene editing, we couldn't mitigate it, we didn't stop it, but it was the case that there had been so many anticipatory ethical conversations that the international community almost immediately said, this is wrong, right? We don't have that with AI, right? So we have investigative reporting that says, or, incident reporting that says a young person committed suicide and we think it might, it's probably because of this chatbot and the parents have all of these logs. And I think instead of having a global conversation that just says, that's wrong and this is why it's wrong, we're stuck in this debate. Is it wrong? Is it the chatbot? Is it the child? Is it the, you know.

And I think that being able to have a thicker conversation about values and technology early on, upstream, all the time, gives us some more societal muscle to be able to, I think I have much more clarity about where we want these tools to fit and sit in our world. And what's happening instead is they get deployed and then we're having the values and ethics conversation after the fact, after something bad happens, after something happens downstream. So I would want thick alignment to be, something that happens early on, it's going to have to happen across the whole deployment chain of AI technologies, but a much richer conversation upstream that's, you know, a bit anticipatory, that includes interdisciplinary and collaborative voices, and that allows us to get out of just a very narrow thinking about whether and when an AI technology is quote unquote aligned.

00:43:28 Caroline
I think the worrying thing is, as you said, it's been landed onto us. People are using these systems in very human ways. Yes. And we could have probably predicted that, right? That people will turn to chatbots to have very deep conversations about things that deeply matter to them, things that are difficult in their lives. And then the effects that these conversations are going to have on them, because it's very human.

So I wonder, how do you feel that now it being landed here, we need these conversations quickly. What's helping us do that now? Do you think we've now got systems? I'm thinking like something like V-Valve from the Collective Intelligence Project, for example. But how can we make sure that we now have these conversations at speed, that thick alignment becomes something that is now part of the journey?

00:44:22 Alondra
Yeah, I think we've got a long way to go because we need quite a few people to be involved in the conversation who aren't right now and who aren't because they are told or made to believe that issues around AI are so complex that they couldn't possibly understand or they couldn't possibly be, expect that they could participate in these conversations. And so I think that's a huge blocker. And I think different from the human genome experience, in the 1990s is that there was, there was a different kind of, I think, ethical orientation to society, which said, it's actually our responsibility to explain, right?

This was research in less commercial products, but there was a different orientation. And I think it's more important than ever for companies. And this doesn't have to be regulation. I mean, this is not, this is about your brand. This is about your status in the marketplace. Do you want to be known as the company that harms or as the company that engages the public around the release of the technologies, right? So what I'm not, you know, we certainly need governance and regulation, but I'm not saying like pass a law and make them, you know, in part because I don't think, particularly coming from a US context, that we're going to get that. But it does mean that we that we, that there are, I mean, I think you mentioned the collective intelligence project. I think there's a lot of projects, but the first step has really got to be, that we might imagine, projects that we might imagine, but the first step has got to be that you belong here, that we all belong in the context of these powerful and transformative technologies and a place of having voice about what they are and where they're going. And I think the fact that these are consumer-facing technologies, even as they're being built into things that we don't necessarily have choice in using, or they're being built into apps and things that we use, mean that there's still some space for... for consumers to, I think, demand more and better.

00:46:37 Caroline
So it seems the missing piece here really is this civic accountability, these mechanisms where people also feel like they've got the voice and the power in order to, you know?

00:46:52 Alondra
The world, as the world of AI governance is very much a world of David and Goliath, right? Like, 5 to 7, however, depending on how you want to count, of the largest multinational companies literally the world has ever known, and sort of everybody else having to sort of deal with what's being kind of shipped out into the world. And, one way of thinking about how you deal with scale is with scale.

There are in the United States alone 330 million people, right? Nearly a billion people in places like India. I mean, they're, and so that's a, so where things like the Collective Intelligence Project or other projects that allow public opinion to scale, public deliberation to be more democratic and scale, for me are places where I see sort of hope and opportunity, where folks can come together in the context of polarization. Obviously, it's a difficult challenge, but to have a collective voice around these things. And I will say, even in the United States, deeply polarized right now, it's not telling you anything you don't know, AI governance remains probably the only bipartisan issue. It's an issue that people are very concerned about, and so I'm not optimistic, but I'm hopeful.

00:48:17 Caroline
That's good to hear that you are hopeful, at least here, because I also feel... So working for an AI ethics institute, what we're often getting is that idea that AI ethics means that we may even be anti-AI, anti-technology, and certainly that we're trying to slow it all down, that we want to slow down innovation.

So what would you say to companies who are all about the race to the bottom, who want to develop and put into the domain as quickly as possible? How would, this thick alignment, these processes and the civic kind of accountability, is there something where you feel, they might say, oh, well, this is just going to slow us down. we don't really want this. What would you say to them about that?

00:49:05 Alondra
I think, sure, a lot of companies are saying that, but I also think, I mean, we've touched on it a little bit, that what they ultimately want, I mean, you know, so I guess this is a moment for how do we think about sort of another way of alignment, alignment of incentives. And even if the incentives, if you're thinking about Venn diagrams, even if just the two corners of the circle are barely touching, right? Like, so you might not have sort of deep alignment and overlap, but companies want adoption of the technologies. And I think as we were speaking earlier, right now you're not going to get the levels of adoption that like really normalize a technology in a society when you have levels 50% or higher of mistrust and distrust in the technology. And so it is, it remains, I think, a fundamental choke point in the success of some of these companies in their products, even as the markets are speculating widely and lots of speculative value is being made around them. I think ultimately, you know, you're going to need the adoption. So I think there's that. I think many companies don't want to be racing to the bottom. And I think if you can place some sort of threshold around the state of play, then everyone can have a baseline that they're playing from and everyone's not sort of just, again, doing this race to the bottom.

So you will hear sometimes quietly from companies that they wish that there were some guardrails and some regulations, some things that would help them be on the same playing field with other companies as opposed to sort of living in the land of kind of shenanigans and hijinks and race to the bottom. So I think there's that as well. And I also would say as a researcher that true innovation, I think a couple of things. I think in democratic societies that true innovation has to include thinking about values and democracy. Like I just, we want to give the title, the classification, innovation, innovative to like lots of different things. And I actually think we should withhold it to think, and use it much more sparingly to include things that actually are enhancing the public good, are offering more opportunities for more people as opposed to just getting more people addicted on social media feeds or something. I mean, like we really need to reframe. both innovation and what value to society is.

And then the last thing I would say in answer to this question is I have, there's an American basketball player named Steph Curry, plays for the Golden State Warriors in my home state of California. And he's like a beast at the three-point line. Like he's just like, he's just like, you know, amazing, just three-point shot, shot or shooter. And so, you know, I have this kind of Steph Curry theory of AI innovation, which is like, if you don't have the constraint of the three-point line, you don't have the majestic athleticism and artistry of Steph Curry, right? That like constraint create compels innovation. And so I also would want to challenge, I think, companies and our colleagues who are working in companies to do innovation that brings that level of creativity and brings that solution set to not only a product, but to big social problems and social issues that really matter for people in their lives.

00:52:51 Caroline
I really love that sound bite and that I find that really exciting, that idea that rather than just thinking about how we're going to get people addicted to our product and make sure they come back for it, is to actually think in different ways. How can public trust help us to roll out the system more?

00:53:09 Alondra
How can it be a necessary, inherent component of innovation?

00:53:12 Caroline
Yeah. So thinking outside the box and allowing themselves to do that. And I guess that's where the interdisciplinarity comes from in again, right? Working with people like yourselves with universities and so on to think that way. So Alondra, there's another part to your Accelerator Fellowship project here. And you're here in Oxford working with a group of incredible people. Can you tell me a little bit more about what you're doing, what this group is working on?

00:53:42 Alondra
Yeah, so this is the second part of the fellowship is the AI Policy and Governance Working Group, which is a multi-sector, multi-perspectival group that I started in 2023 with colleagues from industry. So it's a collaboration with colleagues from Google DeepMind in particular. And I left my work in the White House, and it was very clear that policymakers were going to need guidance around how to think about AI governance issues. It was also very clear that they were getting a lot of contradictory messages. And if we think back two years, there were a lot of the messages where there are warring tribes of AI researchers and they don't agree on anything. And so I think it was very clear to me that in places like Capitol Hill, that policymakers were following these news cycles and sort of like, how do we get expertise around this if everyone's fighting and everyone disagrees and there's all of these kind of warring camps?

And so part of what the working group does is bring together sort of really people from across the spectrum of thinking about AI, I mean, people who very much are concerned about issues of existential risk, as well as people who come from a perspective of thinking about, great possibilities and don't have a kind of strong risk, a lot of thinking about the risk profile, and then people who are worried about current, day kind of harms and risks that we're experiencing already. as well as people who work in academia and AI labs and companies, and also who had been policymakers. And I think this third group is so important to the working group because the work of policymaking is about trade-offs and politics and polar, you know, all of that. And so you're trying to, I think, be creative and think about what policy or recommendations for governance are not only feasible in a kind of legal construction way, but actually pragmatically doable in the context of whatever political madness is happening in any given moment. And so we've taken up lots of issues, algorithmic accountability, how to think about global AI governance.

The conversation that we're having here with the support of the fellowship is going to be about government use of AI in this moment. In which we don't have, maybe the public adoption that I think companies want, but they are very much invested in trying to get government adoption because these are big contracts that often lock in companies, governments for a very long time. And so we're trying to, as a group, think about what are some frameworks and some high-level questions and reframings to offer to local, state, and national governments as they are engaging and making pretty big purchasing decisions about how to use AI.

So there are decisions around the ethics of that, like what should be the ethical protocol, the guardrails for a government as it's using these technologies for things like public services. But also how should we think about, you know, moments, a moment of purchase is a moment of leverage in a transaction for government. And how can, I think what we're trying to think through is are there moments in the, when you're coming to a purchasing decision as a government in which you can include in that issues of public good? So we want, we started talking about vaccines. So if you think about the advanced market commitments that went into vaccines, so the US federal government, I think the UK government did this as well, said, if you can figure out the R&D on this and how to commercialize it, we will guarantee that we will buy X number of vaccines for you, right? So we're going to de-risk for a company.

So imagine, instead of having a model from the 1990s in which, you know, Microsoft or Dell is selling computers or software to a government and sort of locking them into a contract, that, you know, government representatives can sit at the table and say, We're really interested in thinking about working with you on bringing more AI, but we also want to ensure that it has, name a public good outcome, right? More access to more people for... public health information, more access to more people, to certain kinds of educational access, for example. And how do you think about, like, are these opportunities for government to really partner and what AI becomes by sort of making, by having high aspirations for what, how it might be used in government.

So that's just one of the ways, but we're trying to sort of open up a new, I think, framework and kind of choice architecture for thinking about the role of government and this really important moment in which a lot of decisions are going to be made about AI purchases and for that decision not just to be, do I buy A, B, or C, but how can government sort of part of its policy leadership role shaping how those conversations get made and take place.

00:58:51 Caroline
And I can see also there might be, you know, like a conflict here between government also regulating companies and then at the same time, you know, making these purchasing decisions. So that's a really important topic, right?

00:59:04 Alondra
Yeah, and I think there's no easy answers to that. I mean, it's sort of companies are both governing AI and governing with AI in this moment. But one would hope that governments at their best could learn from the learning that they've done around making policy around AI. I mean, part of I think making good use case decisions is also knowing what the tools can do and not do, both their dangers and their possibilities. And I think making prudent decisions as a government and how you want to use them in public services and other kind of, and other aspects of the work of governance.

00:59:44 Caroline
Yeah, I think incredibly important work. And we see it here with some local authorities, you know, commissioners of and health and social care who are making these kind of decisions, but often just

don't really know how do we even decide whether something is a good product. Is it safe? What should we be thinking about? So coming up with their kind of own frameworks here. And so I think it's really important work you're doing here and really exciting.

So it's really great to have you. And tomorrow evening, we'll have a public event on this topic, which will be recorded and available online. So that will be part of your website so people can check in on that too and follow your work and that you're doing with us. Alondra, it's been such a pleasure, you know, talking to you and just benefiting from all your incredible experience and insight and, you know, that thought you've put into everything that you're doing. And so it's such a pleasure to have you with us at the Accelerator Programme, and I'm excited to see what's ahead.

01:00:57 Alondra
Excellent. Thank you so much.

01:00:58 Caroline
Thank you. So to learn more about Professor Nelson's work and other fellows at the Accelerator Fellowship Programme, please come and visit our website, and that's AFP.oxford minus AI ethics.ox.ac.uk. Clunky, but here it is. So this has been the Accelerating AI Ethics, a podcast from the Oxford's Institute for Ethics in AI. If you enjoyed this episode, please subscribe and share, and join us next time as we continue to explore how to build more ethical, inclusive, and imaginative AI futures. Thank you.