I'm Ezra Klein, this is the Ezra Klein Show.
It would be easy, listening to the discourse about AI, to think the government has taken no notice of this technology at all. That there's something happening out here in Silicon Valley and Washington is completely asleep at the switch. It's not quite true though. In 2022, the White House released a more than 70-page document called a Blueprint for an AI Bill of Rights. And the word blueprint there, that is a much more important word in that title than rights. This document is not, for the most part, enforceable at all. These are not rights you can sue to protect. But its release, its creation, was a recognition that at some point soon, the government probably would need to think about creating something enforceable. And so they needed to start thinking about how society thick with AI should look.
What's striking reading the blueprint is that if it wasn't a blueprint, if it actually was enforceable, it would utterly transform how AI has to work and look and function. Not one of the major systems today is even close to conforming to these rules. It's not even clear that if they wanted to, they technically could. And that's what makes this a weird document. Is it a radical piece of work because of what it would demand if implemented? Is it useless because it doesn't really back up its words with power? What is it? And what does it point towards? The process behind it was led by Alondra Nelson, who was a scholar of science and technology, who became a deputy director and then acting director of the Biden Administration's Office of Science and Technology Policy.
So to the extent anybody in government has thought about AI hard and tried to make the sprawling entity that is a federal government develop some kind of consensus opinion on it, it's Nelson. Nelson is now out of the administration. She's a distinguished senior fellow at the Center for Market Progress. And so I asked her to come on the show to talk about what the blueprint did, how she thought about it, what it doesn't do, and how she's thinking about AI now. As always, my email is reclinedshowatnytimes.com.
Alondra Nelson, welcome to the show. Thank you so much, Ezra.
So I want to start with how you and how the government you were acting on behalf thinks about AI itself. Like from the perspective of the government or the public, what is the problem or challenge we're trying to address?
So I would say, first of all, that I'm no longer acting on behalf of the government. So this is a little bit of a retrospective. The office that I worked in is the White House Office of Science and Technology Policy. Its founding statute from the 70s says something like to spur innovation and also to mitigate foreseeable harm. And I think that now 50-year-old statute, I think in some ways at a high level, sums up I think what folks are trying to do with science and technology policy and government and how government thinks about it. It's also the case
that government, and I think particularly the Biden-Harris administration, appreciates that science and technology are supposed to do good things for people in their lives. And so, you know, that innovation should have a kind of mission and the kind of value-based purpose, and that should be for the improvement of people's lives. And I think that's distinctive about how government in this moment is also thinking about these issues. So I'll zone in on that idea of foreseeable harm because I think there are, I mean, there are many, but in this case, two schools of thinking about AI. One is that it has a lot of foreseeable harms. It could be biased. It could be opaque. It could be wrong. And then there's another, which is it has a, this is a sewage and errors

technology. We haven't really dealt with anything like it. The harms are unforeseeable. The technologies, the systems are uninterpretable. And so we're in this place of no one unknowns and unknown unknowns. It makes regulation very hard. Which school are you part of there?
I'm in neither school, actually. You know, I think I am enough of a scholar and a researcher to want more information to think that this is, you know, in some ways an empirical question, a question that we can have more information about before we feel like we have to, I think, plant a flag in either of those camps. I would also say, you know, it's likely the case that it's probably both those things. I mean, there were always harms that we can't foresee or that we can't anticipate use cases that we might have thought about, but didn't consider quite in the right way. So I think across that spectrum, depending on the use case that there are harms that we can anticipate, there are harms we are currently living with, obviously, things like the ways in which facial recognition technologies are actively, have actively harmed Black and Brown communities already. And we have, that's been going on for several years. And then there are,
these are their kind of broader risks. I would say on the latter, you know, we already are living in a time of profound uncertainty with kind of looming risk. And so I'd also want to put the both known and unknown risks that we're thinking about now in the context of the fact that, you know, our lived experience now is that. So I'm thinking of the fact that we've lived for six decades plus with the potential of catastrophic nuclear harm. So that is just something that we that we live with in the day to day. And I'm thinking, of course, of the potential of the existential harm and crisis and catastrophic risk of climate change. And so we live in a world that we're surrounded by risk and to have a kind of new vector for risk, you know, it may be new, but I think the, the conundrum of having to tackle big and large and often unknown situations is not new to human society.
I appreciate those comparisons. And one thing I appreciate about them is that they are comparisons. And I think the human mind often works and the regulator mind, the policymaker mind often works explicitly and more worryingly implicitly through analogy. So I'm curious what analogies you have used or heard in this area and project. And what you find convincing?
I think that the climate change and the nuclear harm are the ones that come immediately to mind. I would say the automated technologies, AI, use the phrase, I think, sweet generous. I'm not sure that it's quite that. I think that there are analogies. I think there's a kind of quilt of analogy that we could put together to help us think about this, not only just the sort of looming potential harm, but also possible solutions. And that's much more what I'm interested in.
So certainly in the nuclear space, you have for a couple of generations now,
activities around non-proliferation. So we know potentially that this thing, these things, these tools could destroy the world. And we are going to work collectively, we're going to work across sectors and globally, and an imperfect way to try to mitigate that damage. And moreover, we have already seen the damage that these tools unleashed can do to the fatal damage that these tools can have. So certainly in the national security space with AI, there's a piece of it that's very much, I think, akin to the potential risks and also a potential strategy or an intervention that comes out of the nuclear non-proliferation space. There is also in the climate space,
I think, the uncertainty of, for example, weather patterns as weather gets more unpredictable, becomes more radical. We might also think of a third analogy here, which would be the pandemic that we're still in some degree living in. And so I think that there's just a profound uncertainty

to life right now and that this is one of them. And it has some pieces of uncertainties that we're familiar with. However, it's also the case that automated technologies are quite literally made by us. And while they can have generative velocity, it's not the case that they necessarily have to be unleashed and unknown to us. That's a choice. And I think one of the things that I hope that we can do coming out of this moment of chat GPT and the like immense possibilities of what automated technologies can bring to the world, both good and bad, is to really think about the choice architecture that's being presented to us or that we're accepting around what might be possible. One thing I notice in those analogies is those are primarily analogies of risk and catastrophe. So COVID, I don't want to take an overly strong position here, but I'm against it. I've been a critic of COVID since the beginning. Nuclear risk, a problem, but also something where a lot of people feel we staunched the capacity of the technology for good because we were so worried about the risk of nuclear proliferation that we could have much more affordable and abundant nuclear energy today if we hadn't been so afraid of the downsides. And of course, climate change, again, I'm going to take a strong position here that I'm against it. How do you think about these analogies in term, because one of the debates in the policy spaces you don't want to staunch technology can they can offer great good potentially, you know, more economic growth, more scientific discovery, more creativity, more empowerment? How do you think about the positive side of the analogies or the concern that the analogies you're offering are too weighted towards a negative? I think the nuclear is actually a quite good example. So, you know, fusion energy has been a promise. You know, there is a possibility of having an exhaustible green energy if we're able to harness fusion energy and fusion research and development.

That effectively is a form of nuclear energy, and that's extremely exciting. We'll have to do lots of things along the way. We'll have to not only just get the science right, which is still lots of work to do there, but, you know, where would you cite the facilities and how do you engage communities around the work and form communities that we are moving into a space in which we were

going to have this tremendous opportunity, but there are also, you know, that they have to learn to think about nuclear energy and nuclear power and different ways. So, all of these, I think, innovations come with both wanting to advance the innovation, but having real limitations.

NAI is one of these. So, there's great potential in science and health. So, when you're thinking about working for President Biden, who wants to reduce cancer death rates by 50% over the next 25 years, you're looking really closely at cancer screening, at radiology, and imaging.

There's clear benefits there. There's clear opportunity there to save lives.

So, that's an incredibly positive application. There's also been incredible work at NASA around using AI. So, folks might be familiar with the recent DART mission that was able to shift the trajectory of an asteroid that might have been, you know, sort of plummeting towards Earth. Artificial intelligence for years has been and remains really central to that work. So, thinking about how you model asteroids, their shape, their speed, their velocity, how fast they're spinning. So, the very future of planetary defense, and indeed, perhaps the very future of our planet, depends on that kind of research and the ability to use that artificial intelligence in that research space to help create action and intervention.

So, there are all sorts of cutting-edge uses for modeling and prediction and more, and that science space as well. There's lots to commend about automated technologies.

One distinction, I think, in the question of automated technologies that comes up in some of the examples you just offered is the difference between these specific machine learning algorithms
and functions that we've already been using can imagine using much more, right, or building a system to figure out protein folding, or building a system to predict the trajectory and shape of an asteroid, or to better read radiology reports. And then the development of these general systems, right, the race to develop what people call artificial general intelligence, but whether or not you even believe we'll get there, to develop these systems that are learning off of a huge corpus of data, are building for themselves correlations and models of how to kind of put all that data together, and are developing capacities that are unexpected and generalizable and not within the system itself oriented any one direction, right, particularly when you're building things that are meant to interact with human beings in a general way, you're creating something that has to be highly generalized, and it also fools people into thinking it's more, well, potentially fools people at least, into thinking it's more of an agent, of a kind of autonomous intelligence, and maybe it is. How do you think about the difference between those more precise, automated, predictive algorithms, and this sort of large learning network equilibrium that seems to increasingly be dominating?
So I would pick out of what you just said, which I think is a nice presentation of the time that we're in, the phrase that you used was, you know, interact with human beings. And I think that, for me, is the central difference. So to the extent that we have our building systems that interact with human beings, I think that we need to have a different value proposition for how we think about this work. So if we're dealing with work that's about people's opportunities, their access to services and resources, their health care, these are things where I think government, industry, academia need to think in different ways about the tools. And so let's be clear, I mean, the tools might, you know, if we think about sort of large scale, generative AI systems, as being, you know, my friend, Suresh Vankanta Subramium often calls them algorithms of algorithms, like, you know, at scale and velocity, you know, in the space of sort of NASA and asteroids,
the stakes for human beings and for interaction with human beings are different.
And so I think what I would want us to sort of anchor on and think about are those spaces, and that if we can build great things, it can't just be the case that we can say it's okay that we can lose control of them, particularly when it has something to do with people's lives.
I think it's a good bridge to the blueprint AI bill of rights. So tell me a bit about the genesis of this document. For those who haven't read it, it's more than 70 pages. It comes out in October of 2022. So the government has been thinking about regulatory framework here. What was that process? So that was a process to create a framework and a resource for thinking about how we might put guardrails around automated systems and automated technology. So we're able, you know, going back to that sort of founding mandate of some government science and technology policy to both move forward with the innovation, you know, as quickly as we can, but also to mitigate harms as best we can. So the AI Bill of Rights lays out a kind of affirmative vision, and it won't surprise anybody listing here what these things are. And I guess to tie back to where we started in the conversation, that if we anchor our values and the best outcomes for human beings, and if we anchor our policies and the sort of value proposition about what technologies are supposed to do and mean in the world for the people that use them, that whether or not we're

talking about AI that we might have used four years ago or AI that will be released in four months, that these assertions should still be true. Systems should be safe. You should have privacy around your data. Algorithms shouldn't be used to discriminate against you.
So that was the process. And I'd say that there are two, and this is I think really important for the generative AI conversation as well, kind of two areas to be thinking about guardrails before deployment when you can do risk assessment, you can use red teaming. For example, you can have public consultation, which is what we certainly tried to do. Can you say quickly what red teaming is for people who are not familiar? Sure. Yeah. So red teaming is, you know, you send your tool to colleagues or your system, and you just have them beat it up. You have them stress test it.
You have them approach it in ways that are adversarial. You try to get them to try to break it. You try to get them to use it at its best and at its worst to anticipate ways that it's not supposed to be used at all to try to figure out how it might break for an effect.
And then to take those insights into improving the tool or the system.
So the second area would be just sort of after deployment. So you can have ongoing risk assessment. You can sort of monitor, try to monitor tools and ongoing ways. And so to your initial question to the work of what government can do, it can really offer an affirmative vision of how the work of doing science and technology of making automated systems of generative AI needs to be ongoing work that never stops and can sort of model that. And, you know, the National Institute for Standards and Technology developed a kind of risk assessment framework that I think is one another way in which government's trying to do this.
So this is framed as a draft Bill of Rights. When the power of the Bill of Rights as it exists and what the metaphor sort of refers to is, if my rights are violated, I can sue.
What is the power or force of this document?
I think a few things. I mean, one is to go back to that original founding document of the United States, which is the Bill of Rights, as you said, and to sort of say that there are things that remain true over time. And that's when we look back to the Bill of Rights. It's not perfect. It's had amendments. It's, you know, our interpretations of it will change over time as society has changed. But there are kind of fundamental things that should be true about American society.
And so while these are not laws, it is a, you know, it's a vision. It's a framework about how laws that we have on the books might be enforced, about how existing rulemaking authorities might be used. And so it's meant to both remind us that there are laws that endure and also to suggest some ways that laws, policy norms that we already have might be moved into the space of new
technologies. And I think there's a broader philosophy here that's important, you know. Technology moves incredibly fast as we've experienced it acutely over the last couple of months, but we don't need to stop everything and reinvent everything every time we have a new technology so that the social compact doesn't change when new technologies emerge.
And that we can pivot back to foundational principles and that those can be embodied and tech policy and then the work of technological development and specific practices like algorithmic risk assessments, like auditing, like red teaming and the like.
But I want to get at the status of what this document is a little bit more deeply because one of the things that is, and we're going to go into it in some detail here in its details, but I could imagine this document coming out as a legislative proposal from the administration. You mentioned that there are laws on the books we could use, that there's regulations on the

books we can use. It's all absolutely true, but to instantiate this as a bill of rights would require quite dramatic new legislation. We would be saying you cannot publish, create, in many cases, train algorithms that don't conform to the standards we're setting out or don't have the process that we are insisting on. And it didn't come out as that. This came out as a sort of framework for discussion, something that companies could adopt voluntarily.
Is your view, like is what you're saying here on this show or in general, Congress should take this up and make it the law, that it should be something you can sue a company if they don't follow, or this is something that is like a good thing for companies to keep in mind as they build their models. How does this move not to vision, but to force, to control? Again, the bill of rights is powerful, not because it's a vision, but because I could sue you if you violate it. Should I be able to sue open AI if GPT-5 doesn't conform? So the audience for the document is just like the participants that led to its creation, is multifaceted. So in the first instance, certainly I would say if I was still working in government, the president has called on Congress to act in these spaces, to protect people's privacy, to move in the space of competition and antitrust. And so there are lots of interesting draft legislation around enforcing algorithmic impact assessments around prohibiting algorithmic discrimination and doing so in ways that continue to make sure that innovation is robust. So sure, there is a piece here that is for legislators, but there's also a piece here that's for developers. And a lot of, as I said, what this document attempts to do is to distill best practices, best use cases that we learned from and discussed with people, developers, with business leaders, with people working in industry. It's also the case that it's for the general public. So it endeavors to make something that I think is often abstract for people a lot more, to bring it down on the ground, to have use cases, including showing folks where there are existing authorities or existing ways that we might think about it. And to go back, I think what I was saying about the broader philosophy is that, and also to go back to the beginning of our conversation, is not only do we live in a world of growing risk in some regard, one might say, but we also live in a world in which there's going to be increasingly more and different and new technologies. I mean, we think about critical and emerging technologies in the policy space and in government. In fact, because innovation is so rich and the innovation cycles are so rich. That means, I think, a different way of having to think about the role of government and the role of policymaking. And it means that it is not to create a new law every time there's a new technology. It is to say, these are the foundational principles and practices, even as the technologies change. We don't have the capacity to create a whole new way of living of American society or government doesn't have a whole new way of imagining American society every time there's a new technology. But what we can do is continue as technologies move fast to anchor and fundamental values and principles.
I'm going to be honest. I don't think that really answers the question really raised by the document, not just by me. So the document has a number of, I think, really profound ideas about legibility. To quote it, it says, you should know that an automated system is being used and understand how and why it contributes to outcomes that impact you. Right now, I would say, and I think this is the consensus view, that even the developers do not understand how and why these systems are often coming to the views they are. So for it to conform to, I think, the plain language reading of this, we would be saying, you have to stop and make these systems interpretable in a way they're currently not interpretable. And if we don't tell them to

do that, then this is not a right at all. So I guess one question is, am I reading that section, right? But another is to rephrase that first question, which is, should that be a law? Do you, Alondra Nelson, think this should be something Congress says, these developers, you have to do this now or you can't release the system because they're not safe? Or is this just something they should do their best on and try to take into account?
All right. So let's see if we can align by thinking about a more specific use case. So let's think about the space of employment and of hiring practices. Let's think about a particular authority, you know, the Equal Employment Opportunity Commission. Their authority is to enforce civil rights around hiring practices and around employment decisions. So that doesn't change when algorithmic tools are used. And it is not the case that developers of algorithmic systems cannot tell us how those systems make decisions. So I want to separate out the sort of speculative cases of tools that have, you know, a trillion parameters or variables and we can't possibly know. And I also, we should come back to that because I don't believe that to be true. So from cases in which vendors are creating algorithms that companies are using to make decisions around employment, we can know how that algorithm was created. And we should be able to, if someone has a claim around discrimination in the space of employment, tell them what algorithm was used. And to the extent that it's not, you know, a trade secret or other proprietary information, give them insight as you would in any other process used for an employment decision about that. I think in a way, I'm not convinced we should separate out the more narrow and more broad systems, right? As they're building these trillion parameter more generalized systems and trying to now offer plugins for them to be in a million different uses. And we know this is coming. We know they're going to, you know, we'll have trading algorithms that are using these systems. And we, you know, can expect these are going to be brought into business strategy and into decision making and so on. If the public, I think, doesn't decide, we are going to enforce eligibility rule. We are going to say that you need to be able to tell us how this ended up making the decision it did, not just say, hey, it's a predictive machine that we train on this corpus of data and it spit out this prediction. But no, we want to know why it did what it did. I mean, these companies are going to have to build them differently. I talk to these people all the time. They don't know why they're getting the answers they're getting. So I think one question I have, whether within this bill or just within your view of the systems, should Congress, should the public say, hey, until you can make these legible, you can't keep building bigger ones and rolling them out. Like we believe for them to be safe and to protect our rights, we need interpretability beyond what you have actually built into it or figured out how to build into it. And we are just going to put that down in the same way. There's a great analogy, I think, in the draft Bill of Rights here to cars and the amount of not just regulations you put on how cars are built, but that putting those regulations down, you guys point out, has actually increased innovation in the auto manufacturing space quite a bit.
But to do that, we made it so you have to do it. You have to have higher fuel standards, etc. Should they have to have interpretability up to a higher point? And if so, what is that point?
The point is, when it's intersecting with human beings and people's opportunities, there are access to resources and the like. So if someone is building AGI and the laboratory or in a laboratory setting, there doesn't need to be legibility where the legibility needs to come and where Congress can act and where rule makers and lawmakers can act is in the space of domains

in which we say certain things cannot be done around human society. So the legibility rule really applies to specific use cases. So car safety is an example. For example, the employment case that I was telling that we talked about, housing discrimination, access to housing, healthcare, access to healthcare resources, and to healthcare services. So let me sort of revisit because I take your point. It's not that we want to separate the narrow and the generative. We want to separate
the use cases that affect people's lived experiences from the ones that I think don't.
And right now, a lot of the generative AI, in addition to the sort of consumer platform that we're using, is still in the space of being transitioned into different tools. And as it gets transitioned into different tools, that generative AI using an employment tool is going to have to abide labor law and civil rights law with regards to employment. And so it is the case exactly as you were saying that developers are going to have to figure out if they're going to want to use these tools how to abide those laws. So one thing you might hear from somebody more optimistic or sanguine
about AI is that a pretty high level of opacity of illegibility in the system is a price we pay for them to actually work. So I think a good canonical example of this is you can feed some of these systems retinal information, right, just they can basically look at eyes. And we don't really understand why they are able to do this and detect gender from looking at retinas, but they seem to be able to do that. And the system can't seem to tell us. We cannot figure out what it is picking up on that is letting it do something we didn't think you could do, which is predict gender from retina. And you might say, okay, who cares about that? But then when you talk about, say, screening patients for certain kinds of cancers, and healthcare is probably a place where you are going to have fairly high levels of regulation, that when you're doing cancer screening, what matters is that the system is as accurate as possible, not that it is able to explain its reasoning. And in fact, if you slow these systems down to try to get them to explain their reasoning, well, then maybe you're not rolling out these detection systems and people are dying because the public sector is making everything move too slowly. How do you think about that kind of trade off? I actually don't disagree with those use cases. And I don't think that you necessarily need the kind of legibility that we've been talking about in that space where you would need it.
And let's pull the thread on the retina case is if it was used for screening for travel, and someone was told that they could not travel, and you couldn't tell them why they couldn't travel. And the person was a citizen and had all these kinds of other protections. So I'm not talking about a kind of known bad actor here. But if somebody's right to travel was being constrained
by the use of a retina tool, and they couldn't tell you why, or they couldn't actually confirm that it was accurate, there should be a legibility rule there. There should be something that law should be able to say about that person's rights. That person should be able to invoke rights to know why their rights are being constrained. And that's different from doing large scale medical testing in which we are looking at retinal scans to tell us something in the research space, or more generally that retinas tell us something about a social variable. It's when those things become use cases that law and government and governance really come to bear in a particular way.
My name is Abdi Latif-Dahir. I'm the East Africa correspondent at The New York Times.
Speaking to someone in their own local language opens up a level of honesty and transparency

that would not be present when I speak to somebody in English. When I come into someone's home and
greet them in Somali or Swahili, you know, like Habari or Habarizeno, it brings you into the room. I understand the culture you're coming from, and I'm speaking to you in the language that you understand, that level of familiarity. I use that to really get deeper into what's going on.
What I'm trying to do is help our readers understand what's happening here in East Africa and see how it plays a role in the bigger picture. New York Times subscribers keep our journalists reporting from across the map to help you understand the issues shaping our world. If you would like to subscribe, you can do that at nytimes.com.
So, the first line of the first principle, which is safe and effective systems in the Bill of Rights, is automated systems should be developed with consultation from diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system. And I think this question of consultation, and I mean, even open AI at this point has released a sort of a policy or a vision document saying, we believe there needs to be open public consultation and how these systems are built. But there is no standard. And I wouldn't say one is outlined here in the bill or not the bill, the framework for what that would look like. So, tell me a bit about your thinking about input. How do you get enough consultation? How do you make sure that consultation is sufficiently representative? And what does it mean for that consultation to be followed? What does it mean for this not to be just democracy washing where Google holds a couple town halls and says, see, we got the public's input, even if we didn't follow it? How do you actually ensure that this consultation is meaningful?
That's the great question. I think that the particular challenge that we face with AI technologies in particular, but I think any space that intersects with policy and expertise is that it becomes increasingly abstract and I think, unfortunately, abstracted from democratic processes. And so, you know, what the blueprint for an AI Bill of Rights was trying to do in part was to grow the democratic constituency around the issue. And you mentioned the open AI case and that they had done
some consultation and called for more. I think that we can see already with the rollout of some chatbots and some of the generative AI tools, ways that a little bit more engagement might have changed,
I think, where we are. So, I think this is a moment of profound kind of political opportunity and opportunity for democracy in part because these tools are becoming consumer facing. And so, we went from as a policymaker, you know, trying to explain to people what automated technologies are and what the implications might be for five years from now and 10 years from now was sometimes
quite challenging. But because these became consumer facing tools, everyone almost immediately, the 100 million users that we know have engaged in using chatbots, have something to say about these technologies. And I think even though we could have hoped for a rollout that was not, let's just put things out in the wild and see what happens, that was a lot more consultative, it is tremendously important that people who are not experts understand that they can have a roll and a voice here. So, I think we're figuring out what that consultation looks like because to me, there is an increasing social need on the one hand growing for the kinds of consultation that we do in other kinds of policymaking. I mean, we don't say to people, unless you know how to build a house, you can't have a voice in policymaking around housing. And similar, we have to think

about how to do that in the science and technology policy space. But on the other hand, the space of expertise is getting a lot smaller. I mean, there are lots of people working in computer science and data science who are not experts in forms of algorithmic AI or forms of generative AI, but we still need to sort of keep those lanes of I think of conversation and understanding open. And so, I think you'll maybe surprised to hear me say I'm actually quite optimistic about this moment, both because you have companies saying we need increasing consultation and also
because it's an opportunity that the general public is really coming to understand what the technologies might mean. I'm always thrilled for anybody to be an optimist on my show about anything because then I can be the pessimist, which is in some ways my most natural space. And let me suggest two things that I worry I see lacking in the public sphere when I talk to policymakers, when I talk to members of Congress and when I think about how this all might be translated. One is self confidence. I think that speaking maybe of Congress primarily, but much of the government, there is a sense that it almost knows what it does not know. Congress is not confident in its own understanding of technical questions. It is quite confident that it could screw technical questions up by misunderstanding them. And as such, I think that there is a tendency to want to be pretty hands off, particularly when a technology is changing rapidly and quickly. And you might say that's actually a good precautionary principle. You might say that is them being wise, but also it means they don't, they try not to interfere too much in things. There's been a lot of talk of regulating social media, but very little has been done as an analogy. And then the other, which is related is consensus. There is very little consensus on even very simple things in Congress, in the government more broadly. And so the idea that there would be enough consensus to act beyond what, you know, open AI or Google voluntarily wants to agree to, that's a big lift. So the technology is moving forward very fast. The public sector moves quite slowly. The White House released it under you, a draft framework for an AI Bill of Rights that has not been taken up into the next phase of, it doesn't become a law, it doesn't become really much of anything yet. And so I think that that's sort of the question. I can imagine how you could have, this could be an amazing moment for democracy. I think it's really profound to say the public should shape the structure and path of these technologies. But I see in many ways a public sector that is too insecure and divided and has lost a sense of its own creativity, such that, you know, I worry it will move much too slowly and will always be, you know, so far behind that it is not shaping this away from the harms that are coming. I think that's right in some degree. I would want to add the following. So I come to Washington as a relative outsider. So I've written about politics. I had been before coming to OSTP, writing a book about the Obama-Biden Office of Science and Technology Policy. So it's not, you know, that I hadn't been thinking about these issues, but I certainly hadn't been in a role in Washington. So I think I came into the role thinking that there was going to be, you know, I think critical of this, what you called self-confidence or not kind of sufficient confidence around tech issues. And I think having left the role, I think that's partly true. But I also know that there's been, particularly over the last decade, incredible efforts to bring more expertise into government and that many representatives and senators have really great teams that are quite technical, great technical capacity and actually know quite a lot about the things that they are trying to legislate on. You know, so then we're really talking about, I think, around the confidence, Ezra, the sort of space of the political theater, which is often

sort of hearings and, you know, or social media. And, you know, somebody might use the wrong word
or the right jargon, but that cannot be, I think, an excuse or justification for not having a democratic process around science and technology policy. So I'd want to say, you know, yes, it would be great to have many more people working on the Hill who felt a lot more confident, but I think around the most advanced science and technology policy. But it's also the case that I personally, as a citizen, don't want to live in a technocracy and I don't want to live with a government in which, in order to have a say about government, about process, about participation, about democracy, you have to have a degree in computer science. And so to come back to what I was saying before, like, we have got to grow a language for talking about sophisticated scientific and technological issues in this moment. And we also have to grow the constituency around them. And one way to do that is to say, you know, you don't have to know the difference between narrow and generative AI. But you can be able to say, you should be able to expect that the automated systems that are used in your life are safe and effective,
that your data has been protected, that a system that's used to make a decision about your access to a real estate mortgage, for example, that you should be able to get an explanation about how that was derived, or that you should be able to talk to somebody about how that was derived.
For example, if the decision doesn't go in your favor and you want more information.
So I want us to move out of a space as a national community, and which only those who can talk, who can use the right words and the right jargon, and that goes from senators and congressmen and women to citizens, are able to sort of have a voice here.
I think this is actually a really important point. I want to call out maybe a dynamic that I'd be curious for your thoughts on. Mark Zuckerberg and other Facebook executives have appeared before Congress now many, many times. And I suspect hundreds, thousands potentially of questions have been asked of them in these long hearings. And I think that to the extent people know of any question ever asked in any of them, it is this one where Orrin Hatch, who was in his 80s at the time, asks, how do you sustain a business model in which users don't pay for your service? And Zuckerberg kind of smirks and says, Senator, we run ads. And you can argue about what Hatch knew or didn't know there, and if he was trying to set something up. But there's this recent TikTok hearing where I think the sort of famed question out of it was where one of the members of Congress asked TikTok CEO if TikTok connects to a home Wi-Fi service. And he was trying to say something else about other devices on the service, but it was a dumb-seeming question in the moment. And there has begun to be this pattern, I think, where you see these hearings. And people are looking for the question that suggests to them Congress or regulators don't have the understanding to regulate these things, which I think is actually also can be fair, can be true. But somehow that always becomes the headline, whereas a lot of the questions are very good. And a lot of the lines of inquiry developed are very good, and they're just not good clips because they're more complex and sustained. And I don't exactly know what to do about this, but I do think there's a meme, there's an intuition that the public sector lacks the technical expertise to be effective here. And I think that's one reason these exchanges always catch fire because they speak to a suspicion that much of the public already has. They then play well or are boosted potentially on the relevant social media services. And as such, something that you have, I think, identified a few times here as being very important, public confidence, is pretty low around the public sector's capacity around technical issues.

So, you know, we live in a meme culture. And so I think that people are already always going to be looking for the memes more than the nuanced and sophisticated questions. And, you know, as I described it as political theater, and, you know, I think that's largely the case. The work of doing technology policy doesn't happen in these hearings. I mean, the hard work happens in conversations with constituents and, you know, meetings with business leaders and lots of other things besides. But clearly this is, you know, I would not say that it's not an important part of political culture. But I think that we don't do ourselves a service to continually engage in a kind of gotcha culture around political hearings. And that, you know, I think all of us really care about the future of democracy. We might have different opinions about the intersection of social media and of generative AI and automated systems with this. But there can certainly be more expertise. We need more people with expertise to be working in the public sector. That's certainly true. But there are core issues about access to effective, safe services for consumers, for students, for patients that are not about using the right word and not about knowing, you know, technology at the sort of highest level.
To stay for a minute on the question of the public sector's technical capabilities, one thing that is true about the government is that it runs an array of unbelievably technically complex systems and institutions and projects. The recent nuclear fusion breakthrough happened at a government lab. If you look at the Defense Department and the kind of technology that they are directly involved in creating, soliciting, buying, operating, running, pushing forward, very, very, very complex stuff. If you look at what DARPA does, very complex work.
Should we have a public option in this way for AI? And I mean this for two reasons. One is that one of the theories of what OpenAI always said it was doing, I'm not sure that's really what they are doing at this point, but what they said they were doing was that in order to understand AI safety, you needed to actually run one of these really capable models so you could experiment on it and run testing and run research and try to understand how to control it. It's also the view of the group over at Anthropic. So you can imagine maybe the government wants to build its own one of these with a lot of money and a lot of manpower so that they have the ability to really understand it at the level of its own guts. And then secondarily, maybe they want it so they can turn it to their own ends, scientific research or maybe it's about trying to work on deliberative democracy, but something where this could be not just a technology used by the private sector, but one used by the public sector. Most of what I've seen so far has been around the idea of regulating what the private sector does, but what about creating internal autonomous public sector AI capacity? Yes, you've hit on something really important and this is work that's been going on in government for more than a year now actually. And just this past January, there was a report released by the National Artificial Intelligence Research Resource Task Force, which comprised of all sorts of amazing researchers from the private sector, from universities, both public, land grant, small universities, folks from different parts of the United States. And it was exactly this issue that this body was asked to address and they made recommendations about this to the White House and to also the National Science Foundation. And the idea here is that is there a public sector, precisely as you say, opportunity to create at scale a way to stand up a research infrastructure that would broaden access to the computational resources, to the data resources, and I would add, put an asterisk here and say to do so in a privacy protecting way that allows academic researchers, government researchers to be able to work in this space. And indeed, I think to go back to some of our earlier conversation and thinking about

the space of innovation, sort of work in this R&D space and this innovation space,
I think to think about some of the tools that to the extent that some of the solutions here around harms, getting that right balance between driving innovation and mitigating harms,
will fall to the public sector because there may not be a market model there that somebody wants to move ahead with in the private sector. Having the raw resources, the compute power, and the data resources to do it are critically important. So there's a proposal on the table that was brought together by a lot of very thoughtful people thinking in this space.
And it's certainly something that we need to pursue. I think a kind of parallel model of this might be something like ARPA-H, which is the Advanced Research Projects Agency for Health, which has just been stood up in the last year. Part of the philosophy there is that there are things that we need to get improved health outcomes at scale and that enable, in order to do that, we're sometimes going to have to do research or move in an innovation space that will never have a kind of viable commercialization or market power in a way that it's going to make particular money or private equity backers or venture capitalists a lot of money. But there still could be a lot of potential for society. I appreciate you bringing up that there are some draft proposals around this. And one of the questions I have about it, and you and I have talked about this previously, is why there isn't more discussion of a positive public vision for these systems. I find it offensive, actually, as a person, that a technology that is potentially this transformative is really just going to be left up to the competitive race between Microsoft, Google, and Meta to make money. That just seems like a crazy way to do things. But it often doesn't seem we have a language and certainly not a kind of rapid legislating capacity around saying, these are the goals we have for this new technology. We don't want this to just evolve in any way. We don't just want to say, hey, whatever can make money here, be that behavioral advertising
manipulation or AI sex robot companions, we actually want to say that we're interested in solving this set of problems. And we're going to put public money behind it. And we're going to make these enterprises or come up with other advanced public purchase commitments. There's a lot of things you can imagine here. But we're really going to put a lot of money here. But we're going to put a lot of money and we're going to wipe out this set of business models. We're just going to say you can't make money this way. We're going to give you a bunch of options where you could become
unfathomably rich solving public problems. And then we'll let some of these private players figure it out. And I'm not saying I have the right goals or the right structure or any of it. But I find it depressing that there isn't more conversation about this kind of thing from the front. I mean, you see it in some areas, I think climate tech is a place where it's probably most prevalent. We actually do now have a government that has put money and is committed to saying, we want these technologies for these purposes. I don't see it here. And so one, I wonder if there is more conversations like this that you've been in that I don't know about. But two,
how you would think about building those kinds of goals or programs?
I think climate change is such a great example because it was not always the case. And I'm going to go back to the phrase I used earlier that there was a democratic constituency or kind of constituency around climate change politics and around resources for mitigating climate change for not only mitigation, but adaptation policies such that people felt engaged from a lot of different sectors in society and the work. And so I think that we are building affirmative vision

around this part of also where we saw this opportunity is in the kind of conversation. I mean, there's been so many debate cycles around generative AI over the last few weeks, but certainly
one of them has been is society kind of the end of civilization looming or is it not? And I have found, I think all of this debate, I think very disempowering. And for us not to be able to say what we want, what we're moving towards, and we're going to harness this sort of strongest ever technological power, this generative AI power and all of the things around it, whether or not it ever gets to artificial general intelligence, that we're not going to use that to build a society and which people can thrive and which, I mean, we should certainly maybe talk a little bit about jobs and work in which we're going to imagine what it means to work and how people can work with technology and tools in ways that we can imagine less exploitative work. I mean, the dream of the 20th century, people having more leisure time, there's all of these kinds of possibilities. And I do think, again, I do not work in the administration any longer, but I do think that there are various efforts going on that taken together really begin to articulate that kind of affirmative vision that you're asking for and that I'm asking for.
Tell me a bit about the role China plays in the political economy of how the government thinks about AI in your experience. Sure. I mean, I think that certainly it's the case that there are great concerns about China working in this space and that sort of innovation that the Chinese Communist Party is sort of driving. So you will hear, I certainly heard lots of concerns about the almost synergistic way in which sort of R&D, the Chinese Communist Party, the Chinese military work together and how that sort of creates potentially at scale a sort of risk to democracy, to democratic values and the like. And so certainly that was a part of a lot of the conversations that I sat in on with regards to artificial intelligence, but lots of other technologies as well. But it's also the case. So the work of the Office of Science and Technology Policy, just to sort of triangulate this a little bit, is also an office that is its job is about sort of helping to have a robust sort of research ecosystem and research enterprise, both for technology and also for basic science. And part of the conversations that we were also having, in my experience, were about how do we both mitigate theft of IP and of intellectual property, theft of technology, while also ensuring that we have the best scientists in the world working in the United States. And quite a lot of that is about immigration and about creating a space, also an affirmative vision for science and technology and innovation in the United States that continues to sort of bring the very best people and draw the very best people to the table. And that means that even as there are, I think, real national security threats with regards to China in the science and technology space, it also is the case that it can't be a justification, shouldn't be for discrimination or xenophobia in that space. And that's, I think, the hard place that the administration finds itself in, as an administration that's deeply committed to expanding innovation, deeply committed to equity issues and equality issues, and also deeply committed to democracy and using lots of creative new tools like export controls and these sorts of things to make sure that there's sort of economic and national security in the United States.
I have more fear, I think, about how this is playing out. Some of that is threaded through what you're saying there. But I hear a lot about dominance and race dynamics that you saw it from the National Security apparatus. They want to be first to powerful AI. When you mentioned the export controls, I mean, one of the main things we've export controlled are the kinds of semiconductors that China would use to train large-scale AI systems. And, I mean,

national security is a very powerful stakeholder in Washington. I think that would almost be an understatement, and they've been ahead on AI for a long time. And it just seems to me that a lot ends up getting justified under the idea that we need to kind of race China into the finish line here. And yet, when I look at China, and I'm not a China expert, and I would like to understand enforcement of these things better than I do, and it always looks like they've been even more aggressive on regulating AI than we have. So they unveiled these new regulations governing internet recommendation algorithms, a lot of which would apply to AI systems. If you want to make a deep fake, you need to have user consent to the images in a deep fake. You can't just do it because it's cool and spread it around on social media. Like, you actually have to have somebody sign off. And now maybe that's just what the American Press is reporting. And in practice, it doesn't work that way. But I worry that we've turned China, particularly inside the government, into this sort of bogeyman that justifies anything we want to do on AI and racing forward really, really fast, even though it's not clear to me. They don't have even more qualms and concerns about the technology than we do. And I want to say that is not to defend the way China, I think what China might do with AI on surveillance is chilling. And I don't want to see them attain AI dominance. But I also don't want a caricatured version of them to be a way that safety ideas get kneecapped in the United States. I think that's right. I take that. My research as a scholar, in part, is in the space of human genetics and the uses of human genetics sort of after the human genome project, where we first start to see the use of large-scale computational power, big data sets in this space. And so the Chinese researchers, Chinese government has been, obviously, a significant player in this space. I mean, there's been New York Times reporting on how genetic technologies have been used as a way to exploit and to surveil the Uyghur population. And so there are real challenges. And to the extent that we can know, I think, what's happening on the ground with regulations and use cases, there's certainly concerns. I think that we can look back across history, and I think to use your phrase that there's always been a boogeyman cast against different kinds of, I think, geopolitical aspirations. But I think I want to take your deepfakes thread and sort of pull that through and say, there's some good work on this already happening in the United States. And while there is indeed some proposed legislation around deepfakes in particular, it's also the case that there have been researchers and organizations moving in this space, proposing ways to do watermarking, to trace the provenance of an image or text. Adobe has been a leader in this space, the partnership for AI, working with some other media companies like the BBC recently released a kind of white paper and proposal for ethical principles around deepfakes and with technologies and techniques for how to do that and how to think about them. So there are ways in which we can both hope for regulation around concerning things like deepfake, but also I think lean on companies, you know, the work that they are doing in this space, working with civil society actors to create some tools to help us navigate and the world of synthetic media. The European Union has also been trying to think through its AI regulation. It has this kind of draft Artificial Intelligence Act, which is a pretty sweeping piece of legislation as I read it. Can you talk at all about how their approach is similar or differs from the one that the White House is considered here? Sure. So this has been a long process, the EU AI Act. I believe it's going to later this month be finalized. And that is a their approach is really based on risk cases and sort of thinking through if an AI system or tool rises to or not a particular level of risk. That could be surveillance or intervention in a person's privacy. It could be around national security issues and the like and

versus kind of low risk uses. So I think for the US approach, you might think about it and that is offering, A, we're still awaiting formal legislation that would be as robust as something like the proposed and forthcoming AI Act in the United States. But you might think of it as being something that the EU side that sort of combines and draws on or sort of that is similar, excuse me, doesn't draw on because what I'm talking about sort of comes after. But the NIST, the National Institute for Standards and Technology, AI risk assessment framework gives tools for thinking about how risky the particular use of a technology might be and ways that you might mitigate that on the one side. On the other hand, it also has values and principles with regards to people's rights that I think is analogous to the blueprint for an AI Bill of Rights as well.
One question I've had about the EU's approach on this is when we are talking about these generative models that end up having a lot of applicability across different domains, right? It's sort of the GPT-4, GPT in the future, maybe five or six system can be applied to a lot of different things. Does separating the regulation out by use case make sense? A lot of people, probably those worried about more kinds of existential or large scale risks, would say what you have to make sure is that the model itself doesn't become dangerous. And so not aggressively
regulating a model because for now it's called a chat bot as opposed to being called a judicial sentencing bot is applying a set of categories that maybe made sense in more specific technologies, but don't make sense here. How do you think about that difference between regulating the models and regulating the use cases? Depending on the context, you might have to do one or both. I mean, so I think we were just talking about national security and a national security space. You might want to be regulating a model and that might be done in a way that's sort of not publicly known or so there might be some you could imagine for geopolitical reasons, intervention around particular models. So that's that regulation per se, when you're talking in that geopolitical space and the way that we might around an act or a law that would be passed around particular use cases. So this is what's true. I mean, I don't use when we began the phrase,
sweet generous, to talk about generative AI. And I think I'm enough of a scholar to think that that's an empirical question and I would want to sort of get more information to think about whether or not that's true. What we can say is true is that automated tools, automated systems, AI, machine learning, large language models, this whole dynamic and world have an extraordinary, what we can say right now is they have an extraordinary broad scope. And I think what your question is really getting at is the fact that the answer to that question about whether or not it's the tool or the use cases or the risk cases or the potential risk, it's sort of E, all of the above. And so I think that's what's distinctive about automated systems. They are both kind of an infrastructure and an enterprise tool. I found it very interesting that one of the first use cases that we will have will have will be generative AI being rolled into effectively Microsoft suite and the Google suite, you know, so tools that we're already using that we usually think of as often as kind of enterprise tools. But it's also will be the case that they'll be used for, I think ultimately, other kinds of systems around, you know, advanced kind of decision making. And that means that we need policy innovation and creativity that can think about what it might mean to regulate a particular tool, but that appreciates that AI and related sort of automated systems and tools will be increasingly woven throughout society and woven and through a lot of the sort of actions and things that we do in the day to day. And so that means that thinking about it as just one thing is insufficient. And that we will need to, I think, have a lot of

policy space and innovation. And so again, you know, this is why something like the Blueprint for AI Bill of Rights is really about not about the tool and is really about the use cases and about allowing people to have exploitation, allowing people to have, you know, an opt out option, to be able to reach a person to not be discriminated against and the use of these technologies.
How do you think, and this goes to the question of being so generous or at least very unusual, how do you think about the existential risk question, the, you know, the finding in surveys that about when you talk to AI researchers, they give a 10% chance if they're able to create generalized intelligence that it could extinguish or severely disempower humanity. A lot of people are afraid that this could be a kind of humanity ending technology.
Do you think there's validity to those fears?
So we already live, I think, as I said, you know, in a world in which the destruction of the human race is not a zero probability. And, you know, the word alignment is used, I think it's a kind of term of art and it's a little jargon. But, you know, I think folks listening should understand that it's often intended to mean aligning AI with human values kind of in a high sense, at a high level, but often through only technological means. And so what we have been talking about and what I've been trying to add, you know, in our conversation today, Ezra, is a different kind of, well, I'll stick with alignment, but an alignment that's about our sort of values as a society and the democratic sort of levers and policies that we can use that are not only about the technology. And so I think the challenge around the very small group of very talented computer scientists and researchers who are saying that we need alignment and alignment can only be other technology is that that choice architecture is saying already that we can only ever think about this with other kinds of technological processes, that we can't ask other questions about should these systems even be developed? How do we ensure that they're safe and effective in a way that allows other people besides scientists and besides the elite group of scientists who are working in this space to have a say about what that means?
Some of the AI alignment conversation is about what's good for civilization and what's good for humankind. How do we have those conversations in a way that actually includes a larger swath of humankind and what that conversation is, and that it is not just a conversation that inherently creates a race amongst technologists as the only possible solution or the only future vision for what American society, planetary society might look like?
You're a scholar of science and how science has worked in the real world and who has been included and excluded. When you look back, we have so many different examples of how science has been
governed for good and for ill, everything from regulated like cars and seatbelts to international bans on certain kinds of things. What do you find to be the most hopeful or interesting or inspiring example, not of the technology, but of a governance or input or some other kind of public structure that emerged around a technology to turn something that could have been destructive into something that was productive? A lot of the research and writing I've done has been about marginalized communities, about African American communities often.
The thing that has been both inspiring and surprising, even though I know it's to be true, is how communities that have been often deeply harmed by technology historically up into the present are sometimes the strongest believers in its possibilities or sometimes the early adopters and the innovators in this space. The last book-length project I worked on was followed the first decade of direct-to-consumer genetic testing. As much as I heard people in the

communities that I was working with say, I have lots of reservations given the history of eugenics, given how quickly the technology is moving in this space and about what this technology can actually really do and mean for my life. But it was also the case that, in the case of this, some of the research I did, you had 60, 70-year-old African Americans who were the early adopters in
2003 and 2004 of direct-to-consumer genetic testing. This is three and four years before you're having the 23 and me spit parties that are being written about in the mainstream press. I guess I would say all of us do believe in the possibilities for technology and for innovation and for what they can do to improve people's lives. But we need to put that front and center. The technology shouldn't be front and center. That belief that technologies are tools that expand opportunity for people, that potentially extend our lives, that make us healthier, I think should be how I would want to propose that we think about what might be the future for generative AI. I think that is a good place to end. It's always our final question.
What are three books you'd recommend to the audience? Three books. Okay. I'm going to offer a pretty new book and two older books. The first is Data-Driven, Truckers, Technology, and the New Workplace Surveillance. It's by Karen Levy. It was just published at the end of last year. I think some of the conversations that we've been having about generative AI and the workplace, how does automation transform the workplace? It's really about how the surveillance automation has been changing the trucking industry, reconfiguring relationships between truckers and their workers. She tells a story about how surveillance is becoming more pronounced and trucking, but also a story about how automation and surveillance are being resisted by truckers. I think we want to think about, as we think about a future vision, an affirmative democratic vision for new technologies, that there's always a role here for individuals and for people to fight back. I've been encouraged, for example, to see that University of Chicago researchers have developed a tool to help combat the mimicry of visual artists' work, for example, already using automated systems. Even when we're thinking about the question of existential risk, we need to be thinking about these questions, imagining that we live in a dynamic space in which there's going to be an ongoing interaction between technological capabilities and what human communities respond
with. Then two classic books, Tim Woo's Master Switch, my former White House colleague, Master Switch, The Rise and Fall of Information Empires. I think we're seeing a kind of instantiation of what Tim describes, which is this ongoing repeated cycle. We're seeing a hype cycle, of course, around some of the AI work, but a cycle from, in his case, he was writing about the emergence of information technology to industries that become empires and consolidate power. I think we're seeing
the same cycle play out with social media and increasingly also with AI. The cycle's not inevitable. We can intervene on the cycle and make sure that there's more space for more people to have a voice and developing AI and automated systems. The last book is Kendred by Octavia Butler. This is a fiction. It's set during the United States Bicentennial. For me, Octavia Butler, of course, is a leading science fiction writer. It's about living with history. It's about technology and democracy, about moving back and forth in time, but still being anchored and living in resilience. Alondra Nelson, thank you very much. Thank you, Ezra.
This episode of The Ezra Clanjus produced by Roger Karma, Kristen Lin, and Jeff Geld. Backchecking by Michelle Harris, mixing by Jeff Geld and Afim Shapiro, original music by Isaac Jones,

audience strategy by Shannon Busta. The executive producer of New York Times' opinion audio is Annie Rose Strosser, and special thanks to Sonia Herrero and Christina Semelowski.