John Villasenor: So, first of all, I would like to say welcome and thank you to Alicia Solow-Niederman and she is a fellow and lecturer at Harvard Law School. Prior to that she held a clerkship at the U.S. District Court for the District of Columbia, and prior to that she spent two years as a fellow in artificial intelligence at UCLA Law. She holds a B.A. from Stanford and a J.D. from Harvard. And before attending law school, worked for three years at the Berkman Klein Center for Internet & Society at Harvard.

Alicia Solow-Niederman: Well, thank you for the opportunity and I'm just sorry it has to be under these circumstances.

John Villasenor: So let me start, Alicia, by saying thank you very much for talking about this important topic that you're talking about today.

Alicia Solow-Niederman: Right, right. As are we all. And so the topic of course is “holding algorithms accountable” and so we'll be talking about algorithms and we'll also be talking about artificial intelligence, commonly referred to as AI. Can you briefly explain the overlap between those two terms because you sometimes see them used interchangeably.

Alicia Solow-Niederman: Yeah. I'm so glad you asked that because I think being clear about terminology is incredibly important, especially when, as I see our task, we're trying to translate across computer science and engineering and law
and policy and different people might be invoking terms in slightly different ways.

Alicia Solow-Ni...: And AI or artificial intelligence, as you say, is especially hard because there is no one definition. In my mind, and I know this is a broader definition than some folks like to use, AI refers to a class of technologies that rely on some kind of automated decision making that's executed by a computer. So something like Google's predictive search function, which is working by parsing all past searches and then recommending, say, how I might want to complete my search based on all the patterns between the words I'm using and the past searches of all past Google users, count as AI.

Alicia Solow-Ni...: Zooming out, and the reason I go there, is because a lot of people might think of Skynet and Terminator and super intelligence, when you think about AI, and there's a reason. I mean, Hollywood has certainly inspired that, but generally speaking there are two categories of artificial intelligence.

Alicia Solow-Ni...: There's what's called general artificial intelligence, which is used to refer to a capacity that is at human level, so reasoning like a person, and at present, that's not in existence. It's—people are developing it. It's speculative when or if we will ever have such a thing.

Alicia Solow-Ni...: And then there's another category of narrow artificial intelligence, which refers to carrying out discrete defined tasks and it's that latter category of narrow artificial intelligence that I'm going to focus on and I'm going to tie into the algorithmic point more specifically in just a moment.

Alicia Solow-Ni...: So I tend to focus on narrow AI and specifically the dominant right now, which is called machine learning or ML, you might hear. I'm not talking about super intelligence or a machine reasoning like a person. Rather, what ML is really seeking to do is to identify patterns that are based on very, very large data sets and basically make predictions by figuring out connections between the datasets, predictions based on the connections it's identifying.

Alicia Solow-Ni...: So where are algorithms in this? Well, I think it's helpful and I started by defining AI because I like to think about AI as the what and algorithms as the how. So say for instance the task we've identified is maybe we have lots of cute pictures of pets, like thousands of pictures of pets and they're exclusive to the cats and dogs for whatever reason. And we want to figure out which ones are cats and which ones are dogs. But maybe, we don't want to take the time to sort them. So we want the AI to tell us which ones are which. And we could do that by trying to develop an algorithm, a
process or procedure, that could correctly classify the images when we give it a new photo that it's never before seen.

Alicia Solow-Ni...: So at a very high level, we might do that by picking a statistical model. When we feed it a training set of images, can accurately sort them. And it would do that by identifying a pattern, for instance, that maybe cats in general tend to have pointier ears than many dogs and sort of smaller snouts. And the idea is to refine the model until the patterns that the training algorithm is identifying, then provide a sufficiently accurate what's called predictive function to apply to new data.

Alicia Solow-Ni...: So in real terms, we're looking for a final product referred to as a working algorithm that can tell us whether a new photo is more like all of the past images of cats that it's seen or all of the past images of dogs that it's seen and then classify the new picture as a cat or a dog.

Alicia Solow-Ni...: And so, that's sort of the dynamic interplay between the concepts of an algorithm and the meta category of both AI and within it the dominant method right now of machine learning. That, of course, I'd say gloss over a ton of technical detail, but I think at a high level ...

John Villasenor: Thank you very much. Thank you very much. Okay, so now move on to the part of the discussion more specific to your work. You've got a forthcoming law review article in the Southern California Law Review and you make the point in that article that algorithms are often portrayed in the broader press as being more autonomous, more than they actually are. Can you explain that?

Alicia Solow-Ni...: Sure. And you're kind to reference my work. So this has tapered off a bit, but what I have in mind is you might see popular news headlines that say something like the AI is tricking its creators by doing this or artificial intelligence is learning.

Alicia Solow-Ni...: And I think that's actually problematic language because as I just gestured at, at least in its current form, AI isn't actually making autonomous choices. The algorithms that power, what we're calling AI today, are driven by human decisions. We have to make choices about what data we're including. So in the cat and dog example, we chose that we only wanted to focus on those two pets. We didn't also want to focus on reptiles and birds. And that goes all the way from the choice of the data to the choice of you know, what statistical function we think might be the best match for the data to how to adjust the model during training.

Alicia Solow-Ni...: And I want to be clear. I think this quaint example is helpful that it's not just salient in the context of pets and things that are light and fun. This really matters because right now for instance, there are algorithms that are
used to determine high-stakes things like the risk that someone charged with a crime might commit a crime in the future. There are a lot of state laws that require the use of that sort of algorithm.

Alicia Solow-Ni...: And on the surface, we can debate whether or not that's a good idea. That's not actually my point here. My point is that we can't measure that target variable. There's no ground truth of we know whether someone's going to commit a crime or not. We don't actually live in Minority Report. So we have to make choices that something like arrest data is a proxy for the risk of future crime commission. So that we think there's an algorithm that can, you know, sufficiently classify whether a given defendant is like other individuals who were rearrested for similar offenses in the past and then make a prediction about that person's risk of committing a crime in the future.

Alicia Solow-Ni...: And my point here, tabling, because I think we'll get to this later, how such processes might be unfair or might introduce bias is that there were human beings all the way down that decision-making path. You had to make the antecedent decision that rearrests are a good measure of recidivism risk, even if there might be different law enforcement resources or attention in some neighborhoods that affects the baseline rate of arrest. And you also have to make all kinds of decisions about you know, what rearrest data should count. What if you arrest someone but then you don't charge them for a crime? Should that enter into the equation? Or if, say the statute on the books changes. So something that was previously criminalized is no longer criminalized. Should that enter the equation?

Alicia Solow-Ni...: And the bottom line is that it's not so simple. It's not just automated, hit the algorithm, it runs, it's some emergent intelligence. It's an algorithm that operates in social context. So I think it's really dangerous to invoke language or even frame it as something that can be segregated from very human choices.

John Villasenor: Thank you very much. I guess the next question, I also have bundled two questions in one, what are some of the ways in which AI has the potential to improve society? And then conversely, what are some of the ways in which AI has the potential to lead to unintended harms?

Alicia Solow-Ni...: Sure. I mean, that's a huge category.

John Villasenor: Yeah. I know. We can talk for days on that.

Alicia Solow-Ni...: It's great. And I'll start with the positive because I'm grateful for the chance to think of the glass as half full these days. So I'm going to focus on machine learning to try to be more precise in answering your question.
Alicia Solow-Ni...: So part of the power is the ability to identify relationships between things that people might not have seen. Machine learning doesn't work from causal reasoning in the same way as the traditional scientific method. It can parse corpuses of data in much more efficient ways.

Alicia Solow-Ni...: So I'm particularly, I think it's interesting to think about ways this might revolutionize the medical sector in particular. For instance, if ML could look at hundreds of thousands of pictures of tumors and figure out new patterns that correlate with a high likelihood of malignancy and allow people distributed all across the globe to have access to medical care that they might not otherwise have. I think that would be really exciting, as long as we can tackle of course, really significant questions of privacy and data security. And as long as we, I think it's just bears noting, submit such medical innovations to the same rigorous testing we would require of anything.

Alicia Solow-Ni...: But putting that aside, I think you can think about a number of medical applications. Others would be discovering new prescription drugs, looking at symptoms of patients who are not formally diagnosed as having the same condition and connecting the dots between seemingly disparate clusters of symptoms. And I think that that's really exciting. I am someone who believes in using technology to reduce inequity where we can, and I see that as one potential place.

Alicia Solow-Ni...: Now as far as harms, I mean, that's something alone I could talk about for far longer I think than even you John might want to hear. I think there are a couple categories of harm that I'd like to try to hit on at a high level. So I want to tease apart harms that might occur in the deployment of an algorithm from harms that might occur in the creation of an algorithm.

Alicia Solow-Ni...: So the first category of deployment is probably more obvious. Picture something like an autonomous vehicle that killed a pedestrian because the onboard software couldn't recognize that she was walking with a bike and the company had disabled the automatic braking function to try to create a smoother ride. If that sounds oddly specific, that's because that's exactly what actually happened according to an NTSB, at least preliminary report regarding the fatal Uber accident in Arizona a while back.

Alicia Solow-Ni...: And you can think of all kinds of real-world physical accidents, problems that might go awry. You know, there's a famous quote, which I'll slightly butcher, but the idea is that when you create a technology, you also create the accident of that technology. And I think that applies with equal force to AI.

Alicia Solow-Ni...: But the second category, which is a bit more nuanced, has to do with the assumptions and the preconditions that are baked into the creation of the
algorithm. So what I mean by that is first you can have technical problems.

Alicia Solow-Ni...: So I've mentioned a couple times when you aim to create an ML algorithm, you have to define the objective function, the goal that you wanted to pursue. And often you're relying on a proxy, like in the example of recidivism and rearrest data and the objective that the product thinks it's pursuing might not actually be what you wanted it to do.

Alicia Solow-Ni...: So I use this example in my paper but think of you know the classic Amelia Bedelia, where she does exactly what you asked her to do, but exactly what you asked her to do might not actually be what you intended.

Alicia Solow-Ni...: There's a whole category of research on this for something, but now picture like an autonomous cleaning robot that is told to clean up everything as quickly as possible but isn't adequately coded to contend with exterior conditions. So it doesn't take into account the fact that like it's knocking over everything in its path. It's just literally cleaning up everything. That example of course involves an autonomous agent but the same problem recurs at all levels of algorithmic creation.

Alicia Solow-Ni...: Getting to the second subcategory though, I realize we're in subcategories of subcategories. But in addition to the sort of like technical assumptions and preconditions, I think the biggest possible category of harm or at least the one that to me is I focus more is the way that any data that's used in machine learning is going to reflect underlying social, political, economic factors.

Alicia Solow-Ni...: And so any time you are using data, which is what we rely on for the dominant methods, you have a huge risk of unfairness or bias creeping in. And once you start to frame things this way, there are all sorts of questions that pop up, in addition to questions of unequal baseline arrest rates. Like I said before.

Alicia Solow-Ni...: I'll just briefly give another example actually inspired by a different project I was fortunate to work on with some computer scientists at UCLA. One problem is there's no mathematical definition of a fair outcome. They're actually over 21 different possible ways you can measure fairness in statistical methods and there is no single policy understanding, to my mind, where people have been fighting over what's fair, what's ethical, for thousands of years.

Alicia Solow-Ni...: So if you tried to implement something like say a statewide risk assessment algorithm as California actually tried to do and it's temporarily—there's a referendum pending in the fall. So it's been stayed, but that's not the point right now. The point right now is that if you try to
fiat something like that at a statewide level and define a fair outcome one way across the board, you can have different population distributions in particular counties, different geological or geographical sub locations within the state.

Alicia Solow-Ni...: But because of the way that the population is distributed, you have different error rates for different subpopulations. So just because you are trying to have what seemed like a fair uniform policy that was invoked as something objective and could be applied uniformly, you end up very disparately impacting people on the ground, without the policy makers at the top ever thinking about that.

Alicia Solow-Ni...: And that's just one example. The bottom line point here is that what's fair is not self-defining. And I think that unless we have a whole lot of dialogue across computer science, law, policymakers and awareness that that problem can creep in, you risk a lot of harm.

John Villasenor: And following up on that question, given the, notwithstanding the kind of inherent challenges in defining what it is to be fair, for example, what are some ways that algorithm designers can help mitigate the biases that are so present? Not only in the algorithms very often, but in the data that serves as the input for the algorithms?

John Villasenor: I mean, there's wide recognition that bias is a problem, a challenge, and knowing that we're not going to get a perfect solution, how can we do better, I guess, is the question.

Alicia Solow-Ni...: Sure. It's a hard question. So let me say a little more about bias specifically since I was just speaking in sort of broader terms about fairness, which is a corollary concept, but just distinct, I think.

Alicia Solow-Ni...: Another related issue, and this is by no means an original thought, is you might have a dataset that consists of only one group but doesn't adequately represent another group and therefore is biased towards that group. So think about, and this is by no means an original thought, there's fantastic research being done on this, but picture something like a facial recognition algorithm that was only trained on Caucasian faces. So it's not accurate when it's presented with images of minorities or there's a sad, horrible story from a few years back of there's a Google search algorithm that for similar reasons identified black individuals as primates.
Alicia Solow-Ni...: And I think it's imperative on designers to think about who is represented and who is not properly represented in their data set and what are the limitations of that dataset? You know, I think there's actually a project to do nutrition labels-style labeling for data sets so that people can be better informed of, is this dataset actually up to par? What are the minimum standards that we need to be holding it to? Because you end up hurting real people when you don't. And it's not just race.

Alicia Solow-Ni...: Just to give another example and be clear here, another problem that's historically occurred is say you have someone at a company who wants to develop an algorithm to identify top performers and predict which new job candidates are going to be likely top performers in the future. Well, if you do that based on your past high performers and it happens as is the case in many companies for many reasons that those past high performers are overwhelmingly male, then you might end up predicting that the male candidates are better than the female candidates because you're picking up on a problem in the underlying data.

John Villasenor: Right, and just to clarify that the reason the past performers were overwhelmingly male, is because the past employees were overwhelming male.

Alicia Solow-Ni...: Correct.

John Villasenor: I mean, it has nothing to do with inherent ability.

Alicia Solow-Ni...: No.

John Villasenor: It's just if 90% of your employees are male, then a disproportionate number of your high performers are going to be.

Alicia Solow-Ni...: Exactly. And so I think recognizing when that happens, if your data set maybe overwhelmingly represents one population and not another population, that's just sort of red alert for you might have a problem when you try to extrapolate beyond that general population.

John Villasenor: Okay. All right. And so, let me turn to another aspect of your work. One very central focus of your work is governance in artificial intelligence. And in your paper, one of the things that you mentioned is the FDA, the Food and Drug Administration, and in relation to areas where there are some lessons that can be drawn. What are some of the ways in which the government process to approve drugs might provide some guideposts to how we as a society might manage algorithms? And what are the limits of this approach?
Absolutely. So the reason I think about the FDA is I was sort of pondering, as I was pondering the question of regulation of artificial intelligence, what are some other areas that we have technocratic domains that require a lot of subject matter expertise, but at least the American governance system has developed ways forward.

So on the surface, I think the FDA analogy is very tempting, at least in the domain of drug approval. It's an agency that's trying to balance the importance of innovation in the form of offering and permitting potentially life-saving treatments to enter the market and a risk to human life. If a drug is dangerous, not testable enough or adulterated.

And let me just say a little bit about how the FDA works, at least in its modern form. So drugs operate through a pre-clearance regime and its ideal form, tabling questions of agency capture or moneyed interests, which are important questions. But the idea is that before a drug can go on the market, the manufacturer has to prove through scientific evidence that it's safe and effective for a narrowly specified use. And the FDA has adapted to manage pretty well, not perfectly, but we do have drug development despite its problems, despite the lag time, etcetera. It's a very technical specialized process.

And so that suggests that one way you might go if you are worried about risk from AI is to require a company to do a parallel pre-clearance for the market. But where I ultimately come down on this is that at least as an overarching solution for all algorithms, I think that drug development is just too different from algorithmic development for this to be a sound way to go.

For instance, drug development takes years and years and requires tons of upfront capital investment and algorithmic development may require massive computer resources and access to a lot of data, but at least in theory it can happen much, much faster. So unless we want to introduce a whole lot more friction into algorithmic development across the board, I just think there's a fundamental mismatch in terms of speed.

And there are a lot of other really important differences I'd like to just tease out. So one of them is once a drug's approved and it's in use by patients, in part because of the safety concerns, in part because of how long the development process takes and part because of the limitations of what molecules can do. You can't easily change the basic composition. That's not true for an algorithm. Companies can update them very frequently and it wouldn't be practical to get re-clearance every time you tweak the algorithm.
Alicia Solow-Ni...: And even more fundamentally, it's crucial that you keep updating the data for the algorithm to ensure that it reflects ground conditions. It doesn't make predictions based on stale data. So you have to keep validating to ensure accuracy. So all of that really ramps up the tension between government pre-clearance and an appropriately dynamic algorithm. And I think that that's irreconcilable.

John Villasenor: Okay. And well, another domain that you identify as a possible model is environmental law. What are the lessons that environmental law can offer in relation to algorithms?

Alicia Solow-Ni...: Sure. So I think in theory, environmental law is actually the most promising analogy for algorithmic governance because it's another field that has to deal with complex systems and make policy interventions, even in the face of a lot of uncertainty.

Alicia Solow-Ni...: So take ecosystem management, that involves waterways, biomes, wildlife and human activity and it requires a lot of active intervention between public and private actors. And the AI ecosystem is similarly complex and critically, to my mind, private companies play a leading role. There, I think, is both the most important takeaway and the potential rub.

Alicia Solow-Ni...: So environmental law, is a domain where collaborative governance and by collaborative governance, roughly speaking, I mean public-private partnerships that don't treat a private actor as an adversary and instead look towards more negotiated, less top-down, less command-and-control approaches to regulation. And that's sometimes been seen as promising in environmental law.

Alicia Solow-Ni...: But what we've learned over time is that there's also a great risk that private companies won't be held adequately accountable for their actions in ways that make it a true public-private partnership. And the reason I say that the potential rub for algorithmic regulation is because we really need to, I don't think we should focus on out on collaborative governance in the abstract. It really depends on what starting conditions we have.

Alicia Solow-Ni...: And at least right now, in AI, I am wary that we can adequately hold private actors to account if their choices aren't in the public interest. And frankly, a lot of that just comes down to money and investments. At least right now, and this is an American-centered approach. I'll just speak to the US. I think we're very far from a place where collaborative governance can be effective, even in, so recently, what's called the National Select Committee on AI, asked for $2 billion in funding for 2021 and that's just a pittance of what private companies are spending and it doesn't all come down to money and of course there's also classified research that we have here in the US.
Alicia Solow-Ni...: But just in terms of gross investment in basic research and development and establishing a strong public voice as like that there is a public force to engage with private actors, I don't think we're there. And just to tie it back to environmental law, I think if we want to move towards a more collaborative approach because there are a lot of reasons, maybe we don't want the government to handle everything when it comes to AI. We want private innovation. We want private actors. We need to think about how to recalibrate the public private balance using failures and deficiencies and another modality like environmental law as a source of lesson.

John Villasenor: Well, thank you very much. And then, that brings me to a kind of a broader question. How do you think that society should decide how much governance is needed? You know, given the different applications where the costs of getting algorithms wrong is so different. For example, if an algorithm is used in a driverless car and it turns out to be unsound, then there can be injuries or fatalities. And so, obviously the cost is extremely high. And the other hand, if an algorithm used in a movie recommendation engine isn't well designed, it's a lot harder to imagine that the government needs to step in and try to make sure that it gets fixed. So how do you make these sorts of decisions about the level of governance as a function of the domain in which the algorithm operates?

Alicia Solow-Ni...: So that's absolutely the million-dollar question. And you're right that it's not one size fits all. I will say I'm hoping in future work to do some more detailed mapping out of how we know which domains are best suited for more or less regulation. So jokingly I'll say, you know, come back to me in a few years and we can have a follow up interview.

Alicia Solow-Ni...: But in all seriousness, for now, what I'll say is that I think there's something really central in the way you phrase the question, that it's something that society should decide. I feel that there should be some amount of social and democratic accountability here. And right now I worry that having so much decision making authority concentrated in private hands, given the private sectors lead in AI, just moves us in a work towards a world where we're sold a cool new product and we're never offered the chance to consider whether we wanted our data used to build it in the first place.

Alicia Solow-Ni...: So I'm thinking of something like Clearview AI, where scraping ostensibly public datasets let the company build what seems to be, according to reporting, at least, a facial recognition tool sold to private companies that effectively takes away our ability to move out the world without being surveilled. And the speed and scale of that and just the ability of a private company to do that, bound only by norms. I mean to be fair, like any company actually could've done this for quite some time. It was just sort
of mutually agreed upon that this was probably not a step we wanted to take.

Alicia Solow-Ni...: But I don't know that any of us consented to our data being used in this way. Just by say, using a social media site and putting our photograph online, and that seems like a really big problem to me, even if I don't know the right overarching governance solution.

Alicia Solow-Ni...: So that's why I mean my preference is to move towards regulation of algorithmic inputs and why I focus so much on data in my homework. I see that as a way that we can start to negotiate questions of public accountability, questions of what value they want to protect in the underlying data, rather than just having a hammer and being too aggressive in AI regulation in a way that actually risks missing the point.

John Villasenor: Thank you very much. And I've got just got one more question. What is your view on the potential for a soft law approaches to algorithm governance and by soft law I'm referring to voluntary frameworks to standards that don't have the force of law.

Alicia Solow-Ni...: So soft law is interesting, because I think it's very appealing in theory and I worry about how effective it can be in practice. As you point out, it doesn't have binding force and that means that it needs to have very strong and compelling normative force in order to bind the people who would subscribe to it. And my concern for algorithmic development is that there's just too much soft law, actually. We're drowning in soft law. I think there's an XKCD comic on jokingly about, Oh, let me create another standard for that.

Alicia Solow-Ni...: And just see what I mean. Think about all the kinds of players involved in algorithmic development. It's not just one data scientist sitting alone in a room who can adhere to some very crisp code of conduct. They're computer scientists. You have data scientists. You have software programmers. You have hardware specialists. You have, dare I say lawyers and business executives whose input might affect the product development, at least as much as the technical specs.

Alicia Solow-Ni...: So already I start to have a lot of questions of whose code of conduct counts and who's going to feel bound by it if it's not by force of law. And I worry even more because soft law seems more likely, to me, to succeed if we're talking about technical standards because it's easier to make objective specifications. Not easy, but easier.

Alicia Solow-Ni...: But algorithms, as I hope I've conveyed, also involve a lot fuzzier tradeoffs around issues like bias or fairness or privacy. And the problem
there is how to give ethical standards of this sort any bite, especially because of the plethora of different standards you might invoke.

Alicia Solow-Ni...: That said, I don't want to come off as hostile to soft law approaches. I'm actually not. I think they could play a necessary role, especially if there were broader efforts, perhaps dare I say with the government in a convening role, not dictating what standards should be, but trying to cultivate that maybe there is an algorithmic professional ethos.

Alicia Solow-Ni...: What I'm talking about is culture shifts within design communities and recognizing that say maybe it's not appropriate to scrape even publicly available data and repurpose it for another context for which the user didn't consent. So I think soft law could be necessary. I don't think it's sufficient. I do think shifting norms is incredibly important and could play a role in allowing us to evolve towards more meaningful governance.

John Villasenor: So, and I'll just ask in closing whether there's any other closing comments that you'd like to offer on top of these very thoughtful observations you've given.

Alicia Solow-Ni...: You know, we've covered so much terrain. It's fun. I guess my one closing comment would be to encourage all of us not to have an allergic reaction to the word regulation, if you hear regulation, and also not to think that regulation is the foe of innovation.

Alicia Solow-Ni...: I started this off by saying how we define words matter a lot and I think it does, and that regulation across the board isn't a bad thing and that innovation across the board isn't a good thing. It's far more nuanced than that and I encourage us to have open minds in thinking, well, wait a minute, are there different tradeoffs we can strike depending on the underlying values we want to realize and not just assuming one thing is automatically good and one thing is automatically bad.

John Villasenor: Okay. Well, thank you very much, Alicia. Very much appreciate this and thanks again for your time and for your perspectives.

Alicia Solow-Ni...: Thank you so much, John.