

algorithmic-fairness

Feb 15, 2026, 5:27 AM

algorithmic-fairness

(0:00) If we collected up all the algorithms in the world, or not just that, what if somehow we (0:06) enumerated every possible algorithm there was? Could we go through them all and label those (0:11) that are interpretable and those that aren't? Well, that's sort of a matter of debate. (0:16) Understanding a model doesn't just require machine learning knowledge, it might also (0:19) require some specialized knowledge about the data set or the domain. Not always, but sometimes.

(0:25) And although we'll cover many different definitions of interpretability during this season and (0:29) explore various angles on what makes something interpretable or uninterpretable, one of my major (0:35) takeaways is that it's a property that's personalized. Something might be interpretable (0:40) for me and not for you or vice versa. But what about ethical? Could we enumerate all the algorithms (0:45) again and go through one by one labeling those that are ethical and those that are not? Binary (0:50) search? Seems pretty ethical to me.

Although what if binary search is employed as one step in some (0:56) sort of virus? Well, again, that's a matter of a hammer can pound a nail or pound a finger. The (1:02) ethics transcend the tool. But those algorithms that make critical decisions that affect people, (1:07) certainly we'd like them to have the property of being ethical.

Interpretability does seem to be (1:12) linked a little bit with algorithmic fairness in that if an algorithm is truly interpretable, (1:17) one can inspect it and perhaps find biases or the absence of biases towards saying that something's (1:23) ethical or not. But it did occur to me recently that to be ethical and to be fair doesn't (1:28) necessarily require interpretability. What if there were some impenetrable black box, which (1:32) through mechanisms we couldn't completely understand, it happened to make quite ethical (1:37) and socially aware choices? There are some other ways we can look at this.

Perhaps, you know, (1:43) we have no knowledge of an algorithm or what goes on inside of it, but we get certain guarantees (1:47) about its fairness. As you can imagine, there's many techniques of anonymization. We'll be covering (1:52) those this season, of course, but there's also the opportunity for some more fundamental or (1:57) theoretical guarantees, kind of the way encryption relies on certain guarantees that computer (2:02) science makes.

One of the many topics I'll cover today is differential privacy. Perhaps using the (2:09) right algorithms and mathematical techniques and things like that, there are ways we can (2:13) ensure fairness and ensure privacy, at least as long as they're baked directly into our design.

(2:24) Welcome to Data Skeptic Interpretability, a podcast about algorithmic fairness, privacy, (2:31) and of course, machine learning.

This week on the show, we've got an interview with Aaron Roth. (2:35) Aaron is a professor at the University of Pennsylvania in the Department of Computer (2:39) and Information Science. He's also a co-author of the book, *The Ethical Algorithm*.

He and I discuss (2:45) socially responsible algorithms, differential privacy, and what promises tools like differential (2:51) privacy can give us. All that and more right after the break. My name is Rita Salam.

I am (3:02) an analyst with Gartner. I specialize in data and analytics. I caught up with Rita recently (3:07) to discuss the upcoming conference in Grapevine, Texas, the Gartner Data and Analytics Summit, (3:13) taking place March 23rd through 26th, 2020.

We expect over 4,500 of the world's leading data (3:21) and analytics leaders. Can you tell us about some of the programs and unique events at the (3:26) conference? One big part of the program is the peer networking experience. And so we have a lot (3:33) of facilitated sessions where we encourage leaders to engage with each other and share best practices.

(3:39) We also have a unique experience that help companies who are looking for maybe a hands-on (3:46) experience to assess different vendor software. You know, you get sort of insight from Gartner, (3:52) which is vendor neutral, of course. We have the analytics and BI bake-off, as well as the data (3:59) science and machine learning bake-off.

To find out more about the Gartner Data and Analytics Summit, (4:04) visit [Gartner.com slash U.S. slash data](http://Gartner.com/slashU.S./slashdata). More details on the full conference schedule (4:10) and how to register, that's [Gartner.com, G-A-R-T-N-E-R dot com slash U.S. slash data](http://Gartner.com/G-A-R-T-N-E-R.com/slashU.S./slashdata). (4:22) My name is Aaron Roth, and I'm a professor of computer science at the University of Pennsylvania.

(4:27) My background is in theoretical computer science, which means that I deal in theorems and proofs (4:34) and definitions more than I deal in code and data. But my particular interests recently have been (4:41) in the ways in which machine learning interacts with social norms. So things like privacy and (4:48) fairness and the way in which machine learning interacts with incentives, how the algorithms (4:54) you deploy can feed back into the ways in which people interact with algorithms.

(4:59) In some ways, it seems that algorithmic fairness is in a very disparate place than these more (5:04) fundamental theoretical ideas. Or do you see them as perhaps more connected? (5:08) I think they're much more connected. So, of course, theoretical computer science is quite broad.

(5:13) But what theoretical computer science takes seriously are definitions and their implications. (5:19) We've seen how useful this can be in the development of cryptography and the development (5:23) of the foundations of machine learning, and especially when the definitions are so important (5:28) to get right. And this is the case when we're talking about important things like privacy (5:32) and fairness.

I think the theoretical approach in which you first think very hard about what (5:38) you actually mean when you say that you want something to be private or you want something (5:42) to be fair, and then carefully think through what the implications of imposing these constraints (5:49) are. It's here where the theoretical approach is especially valuable. (5:53) Privacy is one of those things that I feel like everyone sort of intuitively thinks they (5:57) understand and obviously are for it.

But coming down to a rigorous definition, that's not always (6:02) the easiest thing. Are there ways in which you can define what privacy is? (6:06) Yeah, that's great. And maybe it's useful if I first go through some ways in which you might (6:12) incorrectly think about how to define privacy before finally getting to what I think has become (6:17) a relatively widely accepted definition, which is differential privacy.

(6:22) So for a long time, people didn't think hard about what privacy should mean. And they thought (6:27) about privacy basically through the lens of anonymization or de-identification. So, for (6:33) example, if I wanted to release a medical dataset to the world, what I might have done 20 years ago (6:39) is I might have scrubbed names from the dataset.

If I was careful, maybe I also scrubbed other (6:44) unique identifiers like social security numbers. And then I would have declared the dataset to (6:49) be anonymized and just published it to the world. And this approach is fundamentally broken.

And (6:54) we've known this for a long time, basically because datasets can be cross-referenced with (7:00) other information that's out there. And a surprisingly small number of idiosyncratic (7:05) facts about yourself tend to be enough to uniquely identify you. So maybe the first time this was (7:10) convincingly demonstrated was in 1997, when the state of Massachusetts released a supposedly (7:17) anonymized set of medical records for every state employee publicly to the world.

And this was (7:23) done with the best of intentions to aid in medical research. But Latanya Sweeney, who was a PhD (7:29) student at MIT at the time and is now a professor at Harvard, figured out that although there were (7:34) no names in this dataset and the information that was in this dataset like zip code and age and (7:41) gender alone weren't enough to uniquely identify anyone, in combination they were. And so she (7:48) cross-referenced this dataset with the voter registration polls in Cambridge, Massachusetts, (7:53) and was able to identify the records of Bill Weld, who was the governor at the time, (7:57) and sent them to his desk.

And so anonymization basically doesn't work. And it was things like (8:02) this that forced people to try to think carefully about what they actually might mean by privacy. (8:07) And so maybe here's one attempt that is too strong, but will sort of lead us towards the (8:12) right definition.

So imagine that I wanted to define privacy as the following. Maybe some (8:18) analysis of a dataset is privacy-preserving if after the analysis is done, you know nothing (8:25) new about me compared to before the analysis was done. And that sounds like a pretty good strong (8:32) definition of privacy.

If we could accomplish this, then surely this would be privacy-preserving. (8:37) But it's too strong in the following way. So imagine that you are running some medical study, (8:42) maybe including my data, and you're trying to figure out whether smoking and lung cancer have (8:47) any correlation to one another.

And of course, if you analyze the data, you find that smoking (8:51) and lung cancer are correlated. And I'm some smoker out there. Everyone knows I'm a smoker.

(8:56) I don't try to hide it. And after this study is complete, people now know something new about (9:01) me that they didn't know before. In particular, that I am at higher risk for lung cancer.

And (9:07) this might have concrete harms for me. For example, it might cause my health insurance (9:10) rates to go up. But if we wanted to call this a violation of privacy, there would be two things (9:17) that really went wrong here.

So first, we wouldn't be able to conduct any scientific studies (9:21) at all, because I could tell a similar story with any possible correlation that you might (9:27) discover in the data. And so asking for this as a definition of privacy would just rule out all (9:32) data analysis. But maybe more disturbingly, the story would play out in exactly the same way, (9:37) even if my data had not been included in the data used for the medical study.

Because the thing that (9:43) caused you to learn something new about me was not because you learned something idiosyncratic (9:49) to my data. It was because you learned some fact about the world that smoking and lung (9:53) cancer were correlated. And that just wasn't my secret to keep.

And so differential privacy (9:57) is a very similar definition, but with a little tiny twist that makes it realizable. So differential (10:04) privacy asks you to imagine a world in which the same study is carried out, but without my data (10:11) used anyway. And it asks you to presume that if my data wasn't used at all, then this study shouldn't (10:18) be said to be a violation of my privacy.

That seems rather obvious. Yes, it seems rather obvious. (10:24) So let's call this the ideal world.

Now in the real world, of course, my data is used. But what (10:29) differential privacy asks for is that there should be no way for anyone to tell, substantially better (10:35) than random guessing, whether the study was conducted in the ideal world or whether it was (10:39) conducted in the real world. And if there's no way for people to tell substantially better than (10:44) random guessing, whether we were in the ideal world in which we agreed that I have no privacy (10:48) violation at all, and whether we're in the real world, then maybe we shouldn't think of the study (10:53) that's conducted in the real world as substantially violating my privacy.

And that's differential (10:58) privacy in a nutshell. So I imagine it would be obvious that these techniques are very appealing (11:02) to me. And I love the idea of linking my privacy with problems that are like provably complex in (11:08) the way encryption is pretty provably hard to solve.

And I don't want to put words in your mouth, (11:13) but I imagine, you know, as fellow computer scientists, we can agree that hard problems (11:16) do exist. So it's a nice idea. Let's hitch our wagon to one of those hard problems and say, (11:21) in order to violate my privacy, you have to solve a hard problem.

Cool. Perhaps I have a little (11:25) nagging fear that an especially clever person will come around and find some really novel technique (11:31) to exploit some bit of information and reverse my privacy in a way no one expected. How concerned (11:36) should I be? Yeah, well, that's exactly the power of theoretical computer science.

So the old (11:41) fashioned way of thinking about privacy, this approach of anonymization, that was entirely (11:46) heuristic. And there, your concern is exactly right, that I might think I'm being really clever, (11:53) and there's no way to re-identify anybody. But tomorrow, someone even more clever might come (11:57) around and prove me wrong.

And this is actually how data privacy played out for decades. And it (12:03) was a losing game for the people trying to provide privacy. But differential privacy starts with (12:08) this definition of what privacy means.

And once you have a definition, you can design algorithms (12:14) for which you can mathematically prove that they satisfy differential privacy. So it's this very (12:19) strong guarantee. You no longer have to worry about how clever this hacker is going to be, (12:25) who's going to come around tomorrow, because you've already proven that no matter how clever he is, (12:30) there's just no way for him to tell whether your data was used or not.

(12:34) I see some interesting parallels here with encryption. The world does seem to have accepted (12:39) encryption, by and large, in SSL, even if they don't know what SSL or RSA means. Encryption is (12:44) linked to a computational property.

We're very sure that breaking encryption would be a hard (12:48) problem, and the world seems to accept that. You know, footnote for some things about quantum (12:52) computing, but if we take for granted that new encryption techniques will emerge, will we be (12:57) able to achieve a similar level of trust in interpretability methods amongst the general (13:01) public that they say, yeah, you've got it, that we understand the black box? (13:05) Yeah, that's a great question. So I think part of it will just be familiarity with the techniques, (13:10) because as you say, encryption is also a mathematical topic.

But I think we've gotten (13:15) to the point where people trust encryption, even though they don't necessarily understand (13:20) the nuts and bolts of how it works, in part because encryption has become a widely deployed (13:25) technology that you use every day when you, for example, buy things on the internet. (13:30) And we're not there yet with differential privacy, but we're getting there. So for many years, (13:34) differential privacy was a purely theoretical topic.

People like me would write mathematical (13:39) papers for the small community of other people like me, and these papers would basically just (13:45) prove theorems and think about tradeoffs and algorithms. But in the last, I'd say, four or (13:49) five years, differential privacy has become an actual real technology that started to be widely (13:54) deployed. So for example, if you have an iPhone in your pocket right now, it's reporting usage (14:00) statistics back to the mothership in Cupertino using the protections of differential privacy.

(14:05) Google Chrome does similar things. And the real moonshot for differential privacy is going to come (14:09) just next year, when the 2020 US census is conducted, because the census has committed (14:16) to releasing all statistical products that result from the 2020 census using the protections of (14:21) privacy. So I think that although it's a new technology, as it becomes more broadly adopted, (14:28) it will gain consumer trust in the same way that encryption has, even though as a mathematical (14:34) topic, it might continue to be the kind of thing that only experts really understand the details of, (14:40) again, just like encryption is.

So I feel pretty confident that I could make a pitch to a friend, (14:46) a non-technical friend, over dinner or drinks as to why they should trust encryption. Could you (14:51) convince them that differential privacy is a trustworthy technique? I think the way to think (14:56) about differential privacy is as a form of plausible deniability. What it means is anyone (15:02) who looks at the outcome of some data analysis task, some deployed machine learning algorithm, (15:08) some private data release, and thinks they've concluded something about you, can reasonably (15:15) be told that they are wrong and that you didn't participate in the data set at all, or you did (15:21) and your data was different.

There are simple examples of algorithms that achieve differential (15:25) privacy that make this more transparent that we can talk about if you want. So maybe this is the (15:31) simplest example of how you would actually achieve differential privacy, which otherwise might seem (15:36) a bit exotic. So suppose I wanted to conduct some poll of the residents of Philadelphia and ask them (15:44) something embarrassing.

Maybe I want to know how many people in Philadelphia have had an affair. (15:48) So one way I could conduct that poll, the standard way, is I could call up people on the phone, (15:53) I'd get a representative sample, and I would just ask them, you know, have you had an affair in (15:57) your marriage? And I would write down the answer, maybe in an Excel spreadsheet. And in the end, (16:02) I'd try to compute some statistics, the average, maybe some confidence intervals, and be done with (16:08) it.

Now I've never been to Philadelphia. Are these Philadelphians especially known for honesty? (16:13) Exactly, this is the problem. So as you correctly noticed, people might be reluctant to answer this (16:19) question, because it's compromising information.

And even if they personally trust me, they might (16:25) worry that this Excel spreadsheet that I'm compiling with this very sensitive question about (16:30) them will get into the wrong hands. You know, you could imagine that it's lost or stolen, or that (16:35) it's subpoenaed in a divorce proceedings or something. And so they might have really good (16:39) reasons to not want to answer my question honestly.

But here's another way I could conduct (16:43) the poll. I could call people up again, and I could say, have you ever had an affair? But then (16:48) I'd say, wait, wait, wait, don't tell me the answer just yet. What I want you to do is I want (16:52) you to flip a coin.

Okay, don't tell me how it comes up. If the coin comes up heads, I want you (16:58) to answer my question truthfully. But if the coin comes up tails, I want you to just give me a random (17:04) answer.

Flip the coin again and tell me yes if it's heads and tails if it's no. So I tell people (17:09) to answer in this randomized way, and I can still write down their answers. Now people have a very (17:14) strong form of plausible deniability.

I suppose in a divorce proceedings that my records are subpoenaed (17:19) and the lawyer says, well, says right here in the Excel spreadsheet that you answered the question, (17:25) have you ever had an affair? Yes. Well, you can now quite plausibly say that indeed you answered (17:31) the question that way, but it wasn't because you'd had an affair. It was because that's what (17:35) the coin flips told you to do.

That might well have been the case. So if I gather the data in (17:40) this way, then every single person in the data set has a very strong form of plausible deniability, (17:45) and I as the researcher don't have any strong signal about the answer that any particular person (17:52) really should have provided. But that's okay because my interest in this data set in any case (17:57) wasn't in the behaviors of any particular person in my sample.

My interest was in the aggregate. (18:02) What was the rate at which affairs are conducted in Philadelphia? And it turns out that even though (18:08) each individual response is very noisy, if I only want to know this aggregate, then I can still (18:13) calculate it extremely accurately because I know the noise distribution. And when I average over (18:19) a large number of people, I can subtract it out.

This is sort of a version of the law of large (18:24) numbers. And so this lets me get at statistical facts that I wanted to learn without learning very (18:30) much about individual people. And this is one way to get differential privacy in a way that I think (18:37) makes this guarantee of plausible deniability sort of visceral and easily understandable.

(18:42) So the coin toss idea is really interesting and kind of especially dramatic for me in the sense (18:47) that there's a 50-50 chance here. So the privacy enthusiast in me really applauds that. It seems (18:54) like that gives us a good benchmark of protection.

But the statistician in me is a little bit more (18:59) reserved here and worried about kind of like a noise threshold. Could there be very rare conditions (19:05) or data points, you know, like some medical condition that we would really like to sample (19:09) it with a lot of precision because we don't want to underserve that person, but they're going to (19:13) get washed away in this 50-50 noise threshold. Ultimately, my question is, what do you suppose an (19:19) information theorist might have to say about all this? Yeah, for sure.

And one point that we make (19:24) over and over again in our book is that none of these things like privacy or fairness that we (19:28) might want are going to come for free. So differential privacy by design makes it impossible (19:34) to determine things about a data set that are really facts about just one person. And by extension, (19:41) it makes it harder to tease out facts that depend on just a small number of people.

And so it is true (19:48) that just about any statistical data analysis task, and when I say statistical, I mean any kind (19:53) of data analysis task whose right answers are a function of the distribution, not of particular (19:59) individuals. Any such task can be conducted with the protections of differential privacy, (20:04) but it comes at a cost. And the cost is usually that you need more data to carry out the same (20:10) analysis to the same degree of accuracy with privacy protections than you would need without (20:15) privacy protections.

So you had mentioned that the 2020 census is going to employ some of these (20:20) differential privacy techniques. Could you elaborate a bit on what's going to be changing there? (20:25) That's right. So 2020 will be the first year in which this is done, and it's exciting.

We'll see (20:30) how it works out. But the census collects lots of information about every single American citizen (20:37) and some non-citizens. That's actually one of the sensitive points.

And it releases literally (20:43) billions of numbers, tables of statistics that contain billions of summary statistics about the (20:48) American population. And it's required by law to protect the privacy of American citizens, but the (20:54) law doesn't say exactly what that means. So in previous decades, census used basically heuristic, (21:01) ad hoc techniques to try to offer some form of privacy without spelling out exactly what (21:07) privacy means.

They tried to basically randomly swap people between neighborhoods before they (21:14) computed these statistics. But it turns out, maybe unsurprisingly, that these heuristic techniques (21:20) don't actually work very well. And John Abowd, who's the chief scientist at the census right now, (21:26) led a study that showed that actually using database reconstruction attacks that had (21:31) appeared in the differential privacy literature on these heuristically anonymized tables of census (21:36) statistics, you could actually reconstruct a large fraction of the raw data that the census (21:43) collected that these methods were intended to hide.

And so he decided, I think quite reasonably so, (21:49) that these heuristic techniques weren't actually compatible with the census's mandate to protect (21:54) privacy. And it's in the midst of deciding exactly how to compute these statistics with the guarantees (22:01) of differential privacy, which will involve perturbation with noise. When I say it's in the (22:07) midst of deciding exactly how to do it, I don't mean that they're still trying to come up with (22:11) the algorithms.

We know how to do these things algorithmically, but there's a policy decision (22:16) to be made. Because differential privacy is a definition that comes with a parameter. Remember, (22:22) what it promises is that nobody can tell the difference between the ideal world in which (22:26) your data is not included in the data set and the real world, substantially better than random (22:31) guessing.

So what does this word substantially mean? Well, you can quantify that and you can (22:36) attach a number to it. And this gives you a knob that lets you, in a quantifiable way, (22:42) trade off privacy. You can ask for more privacy or less privacy with accuracy of the statistics.

(22:49) In general, if you ask for more privacy, you're going to get less accuracy and vice versa. (22:52) So I love the framework of a knob that's something that someone can control. (22:57) I'm wondering who should control it.

Is this something that requires an understanding of (23:01) machine learning and some sort of data scientist should be tuning it like a hyperparameter? Or (23:06) maybe somebody who has a history at the census who has some context for it? Or I don't know, (23:10) maybe it can be a more general setting, but it feels like a hyperparameter to me. (23:14) How do you envision it being tuned? Well, I think it's less of a hyperparameter (23:18) and more of a policy decision because there is the algorithmic optimization question. Once you fix (23:25) a particular privacy parameter, the question is how do you actually design algorithms that satisfy (23:30) that level of privacy while at the same time being as accurate as possible? So that question is (23:35) definitely an algorithmic question that mathematicians and algorithm designers and statisticians (23:40) should be working on.

But once that question is solved, there remains the policy question of how (23:46) we should weigh privacy protections with the accuracy of our statistical products. Both privacy (23:52) and accuracy of statistical products are different kinds of goods and different stakeholders care (23:58) about them. And at some level, they are sort of fundamentally at odds with one another, right? (24:03) You cannot ask for more accurate statistics about more quantities without eventually giving up on (24:09) privacy.

So that's not a mathematical question to figure out how we want as a society to trade off (24:15) these different things. That's a policy question that different stakeholders will have different (24:20) positions on. So somebody needs to make the decision.

I don't know exactly who should be (24:25) making the decision, but it's a policy question, not a mathematical question. (24:29) Yeah, that's very interesting. I'm reminded of a thought experiment I've had off and on here.

(24:34) What if I found out there was a cache of medical records released to the public? But then what if (24:40) the next sentence I learned that it was from ancient Roman citizens? I somehow don't have the (24:45) same sympathy for them. I'm not exactly sure where we have the half-life of privacy here, but do you (24:50) have any thoughts on where we draw those lines? You know, I don't know whether my thoughts are (24:55) particularly better than anyone else's thoughts on that question. But I do think it's reasonable (24:59) that we should have fewer privacy demands for the data about ancient Roman citizens.

And I think the (25:08) Census also takes this position. So there is exact data that's been made available from, I believe, (25:14) the 1940 Census. Of course, there's people from the 1940 Census who might still be alive even.

But (25:19) at some point, I think it's reasonable to decide that data is old enough, and maybe the people for (25:27) whom that data was sensitive have now passed away, that it can be made available. And data of that (25:32) sort is quite valuable because it lets you design these algorithms. When you're in the process of (25:38) designing and tuning an algorithm, it's very

helpful to have access to the data of the same (25:43) sort that you're going to ultimately need to run the algorithm on.

And that can actually be quite (25:48) difficult if your only source of data requires privacy protections, because it would sort of mean (25:53) that the actual algorithm design process itself could only access the data in a privacy-preserving (25:57) way. So I think it's reasonable and valuable to think of privacy as something that might be a right (26:03) that degrades with time. Maybe the older the data is, the less expectation we have for privacy, (26:09) especially when this degradation happens not at the scale of weeks or years, but of lifetimes.

(26:16) So computer science people don't swear any sort of Hippocratic Oath. There's no Turing's Oath. (26:21) And I think that's right.

People choose to go work on the next version of Pac-Man. (26:25) There's really no ethical concerns there whatsoever. I imagine if I looked into it, (26:29) someone working on like really deep R&D pharma stuff, they might actually end up swearing the (26:33) Hippocratic Oath.

But in general, for machine learning people, there is no such oath. In the (26:38) absence of one, do you have any best practices or things you think that a practitioner needs to bring (26:44) to the table? To what degree does that technician need to be an advocate for these sorts of techniques (26:49) in their organization? Yeah, so that's a good question. And as you say, there's a wide range of (26:54) roles.

But I think we are at the point where ordinary software engineers are making what are (27:01) in effect important policy decisions without even knowing it, right? If you're someone who's in (27:06) charge of helping to design, for example, the Facebook news feed, then the minor changes that (27:12) you make, in principle, can have large scale effects on, for example, the nature of our (27:18) political discourse. So I do think that it's important that people who are working on products (27:23) that affect many, many people be cognizant of the unintended side effects that their choices (27:30) might make. And I think one starting point is just awareness that there is an emerging science and (27:36) set of technologies designed to both think about what those unintended side effects might be, (27:42) and provide concrete mitigations for them.

So we've been talking about differential privacy, (27:48) people should know that differential privacy is out there. Another thing that's been getting (27:51) lots of press lately has been algorithmic unfairness. I think it would be good for (27:57) engineers to be both aware of the potential for out-of-the-box standard machine learning techniques (28:04) to result in outcomes that seem unfair in various ways, and that there's a large community of people (28:12) and a growing body of science that leads to concrete interventions that can be used to (28:18) prevent some of these harms.

When thinking about what it means for an algorithm to be unfair, (28:24) at some level, I start to wonder, well, you know, we can express those algorithms in terms of (28:29) Turing machines or

assembly code or some higher level source code, where does the unfairness (28:33) actually come in? Yeah, that's a great question. And I would hasten to add that I think unlike (28:38) in privacy, where there's this one definition that many people agree upon, people don't really agree (28:44) on what unfairness is. But maybe I can address your first point, which is to say, an algorithm (28:50) is some computational process, it's a human artifact, right? Does it make sense to talk about (28:56) algorithmic unfairness? For example, another human artifact that's a useful tool is a hammer.

(29:01) And I could use a hammer to do unethical things, I could go around whacking people on the head with (29:05) hammers, but nobody would make the mistake of ascribing the moral failing to the hammer, right? (29:11) We wouldn't talk about ethical hammer design, it would clearly be my moral failing if I, as a user (29:17) of a hammer, went around whacking people on the head with hammers. But, you know, algorithms that (29:21) result from the machine learning pipeline are different, because many of the bad effects that we (29:28) see from machine learning algorithms, and maybe just to be concrete, we can talk about what has (29:35) maybe become the most well-known example, which was unfairness in recidivism prediction that (29:42) ProPublica uncovered a number of years ago. So, just for background, in many states, (29:47) actually, including in Pennsylvania, when judges make bail and parole decisions, they are (29:54) actually given, as one of the pieces of information, the output of a prediction.