AI Ethical Risks with Reid Blackman

Jennifer Cohen and Reid Blackman

Jennifer Cohen 00:04
Welcome back to Voices and Bioethics. I'm Jennifer Cohen, when it's my pleasure to welcome Reed Blackman to the podcast. Reed, thank you so much for joining us today. That's my pleasure. Thanks for having me. Reed Blackman is the author of ethical machines your concise guide to totally unbiased, transparent and respectful AI published by Harvard Business Review, press and due out in July of this year. He's the founder and CEO of virtue and ethical risk consultancy. He's also a senior adviser to the Deloitte AI Institute previously served on Ernst and Young's AI advisory board, and volunteers as the chief ethics officer to the nonprofit government Blockchain Association. Previously, Reed was a professor of philosophy at Colgate University and the University of North Carolina Chapel Hill. He also founded a fireworks wholesaling company and was once a flying trapeze instructor.

Reid Blackman 00:57
Read on your very helpful website, Reid Blackman.com, you provide an AI ethics Crash Course. And you start out by saying that the AI does one thing at its core, it learns by example, that sounds pretty benign. But it turns out that it can lead to a lot of ethics issues. Why is that?

Jennifer Cohen 01:17
So there's a number of things to say here. So it's both that it learns by example, and the way in which it learns by example, that creates a variety of problems. So I like to highlight there's four main ethical buckets of risk when it comes to AI. So learns, by example, the most obvious thing that people talk about is that it can learn from discriminatory or biased examples. So if, for instance, the examples are, let's say, from the criminal justice history, or perhaps examples about who spent what kind of money on healthcare in the past, then it can learn to discriminate against most obviously, Black people, for instance, not giving as much care to Black people under diagnosing them, giving them higher risk reading scores for criminal justice purposes, and so on, and so forth. So I can break that down if you like. But the basics is, if you're learning by example, and you're learning from examples that reflect certain kinds of historical biases or discrimination, then you're going to get bias or discriminatory AI. There are other ways of getting by discriminatory AI. But that's the go to, then there's the way in which it learns from the examples. So what it does is it pours through those examples, you know, slightly fancier word, for example, in this context is data. So it pours through 1000s, and 1000s, and 1000s of data points. And what it's doing is, quote unquote, recognizing phenomenally complex mathematical patterns among all those variables. And so those patterns are so complex that they defy human understanding. So the way in which it learns from those examples is by finding or discovering alleged patterns in the data. And those patterns we can't explain because they're too mathematically complex. And so now you have the ethical risk of unexplainable outputs. So first bucket of ethical risk was discriminatory outputs. Second is unexplainable outputs. Third bucket is privacy violations. And here we've got really two places to look, one is the examples that you use, the data that you use, could be ill gotten data, so data that you have, by virtue of violating people's privacy. Or what do you do with machine learning, you're making certain kinds of inferences, or predictions, which is to say, you're coming to know new

things, the knowledge of which made themselves constitute violations of privacy. So those are the three buckets and the fourth bucket is sort of a grab bag of use case specific ethical risks. There's all sorts of ways you could use AI, that are going to come along with ethical risks that don't relate to the first three that I mentioned. Okay, so hopefully, we'll be able to unpack a lot of that very complex answer. So do I have it right? If I frame it this way, that when algorithms are resulting in these biased results, or in a worst-case scenario, physical or mental harm to someone, it's not so much that it's a technical bug, or something, some kind of computer programming mistake, it's these larger ethical issues that you've already raised around the inputs and the way the data is being manipulated.

Reid Blackman 04:20

Okay, so I wouldn't characterize it as a technical issue, I would say that there are a number of, if you like, tactical decisions, and behind those tactical decisions are ethical assumptions. Maybe that's the way to put it. Right. So let me just give you some examples. Because one thing that people will do in this context, which I find relatively frustrating is to just say things like we need representative data sets, we don't have sufficiently representative data sets. And what they have in mind there is if you have certain subpopulations that are underrepresented in the data relative to other populations. You're going to get it to be trained really well on those majority populations and really poorly on those smaller subsets. So let me take it an example outside of the healthcare world, if you are doing facial recognition software, and you're training it up, you get a bunch of training data of people's faces. And you have, you know, 90 percent of the faces you have are, say, white men. And let's just throw a number out there point 5 percent of them are Black women, it's going to learn a lot about the faces of white men and very little about the faces of Black women, and so be much better at recognizing the faces of white men and Black women. You can do the same thing with diagnosing diseases, you know, making predictions about when someone will develop a disease, so on and so forth. That's a case in which there's under representation of a certain sub population, and that can lead to discriminatory outputs. But there's other kinds of decisions that data scientists make that can lead to those discriminatory outputs. Let's take for example, one thing that you're doing is, so your AI model has a certain set of outputs, and there's going to be a threshold, it makes a prediction about the probability of someone having a disease from zero to one. So you know, if it's point one, three, that's a 13 percent chance, if it's point nine, five, it's a 95% chance. And you've got to set the data scientist sets some threshold above or below which someone is positively diagnosed or not diagnosed as, say, having a certain disease, or in a broader context, being credit worthy, or mortgage worthy or chicken insurance premium of a certain sort. And where you set that threshold, is going to have differential impacts across different subpopulations. You know, obviously Black, white, Asian, then you've got things like Black men, Black women, white women, white men. So where you set that threshold will have differential impacts. And then there's a question about whether that differential impact is ethically acceptable, or if it instead is discriminatory. So that's a threshold issue, not necessarily the training data it. Let's take one more example, one thing that you might be trying to do with AI is, you want to maximize something. So you've got all this data about, let's say you're doing kidney transplants, and you want to make a prediction about who's going to live the longest all else equal, if they receive this kidney, the idea being and here's an ethically laudable goal, you want to maximize years saved with your kidney transplants, all else equal, you don't want to give it to the 95 year old instead of the 18 year old because the 18 year old is gonna get a lot more use a lot more years out of that kidney than the 95 year old. Okay. So you set as your this is the technical term, your objective function for your AI, I want to maximize quantitatively years saved. And now tell me, given all this training data, who should I give it to? Now, who should you give it to? Well, it turns out, at least in America, you only give it to mostly white people. And that's because white people, on average, live longer than Black people. So the problem is not that you don't sufficiently represent Black people in your training data or something like that. The problem is that your objective function has implicit in it, the result that you're going to favor white people or Black people. So now, what you might need to do, among other things, is you might need to change your objective

function, maybe maximizing years saved as ethically laudable as it seems from the start, is not the goal that we ought to have.

Jennifer Cohen 08:22
Okay, so when you say on your website, organizations don't understand the real sources of AI's ethical risks. This is the type of thing you're talking about setting the threshold levels, changing your objective function. Do I have that right?

Reid Blackman 08:36
Yeah, that's fair. I mean, there are lots of so people talk about biased AI. And we're black box algorithms. And they make it sound as though there's a problem, you know, not sufficiently representative data sets. But right, there are multiple sources, and I've only named three. So there's under-representation, there's your objective function, there's where you set your threshold. There's how you weight the input variables. So for instance, let's just take a case of insurance, the state driving insurance you might put into there, you might think it's relevant both age and education level, but you think that their age should count more than their driving level in the ultimate determination of what their insurance premiums ought to be. So how you weigh your input variables is going to have an impact on what your outputs are. And thus, the way in which it's differentially distributed across various populations. And this could well result in discriminatory outputs.

Jennifer Cohen 09:30
Fascinating. And you use the term that I've seen come up a lot black box AI, can you describe what that is?

Reid Blackman 09:37
Yeah, so I mentioned this at the outset. So again, let's take a non healthcare example because it's really simple. And if there's a case that everyone is familiar with, you take your photo software that recognizes it said a picture of your dog. So you upload a picture of your dog Pepe, actually you upload lots of pictures of your dog and you tell the software, this is Pepe, you know you type in next to the photo, Pepe and it learns, quote unquote learns what Pepe looks like. So that when you upload that, you know, 100th photo 1001th photo of Pepe, the software says that's Pepe. And if it's not Pepe, then it says not Pepe. What is it looking at now you and I are looking at Pepe's, eyes, nose, mouth, tail, etc, etc. The AI is not looking at that it doesn't look at macro features like that it's looking at the picture at the pixel level. Of course, we don't look at the pixel level, we look at it holistically, but the software looks at it at the pixel level. And so it's analyzing those pixels, and the mathematical relations among those pixels in dozens, hundreds 1000s of photos of Pepe and it's looking for what you might call the Pepe pattern, what's the mathematical pattern of pixels arranged thus and so such that this is Pepe. Now whatever that equation is, whatever that mathematical pattern is, it's too complex for us to get you and I unless you've got some phenomenal hidden genius of which I'm unaware. So we get the language in which it speaking, we get the math language, but the pattern is too complex to hold on our heads. Now, in that case, we don't really care when it says this is Pepe, it gets it right, we're happy because all we care about is accuracy in that sort of instance. But if you say something like does this person have cancer, and the model says yes, but you don't know how it arrived at that output, you might think that's doesn't feel so reliable, it looks like on the face of it, we need to understand how the AI came to make this substantive prediction about this person. So when you don't have explanations when you don't understand why the model arrived at this output that it did, that is a black box. I'll mention one more distinction that I think is really interesting and often overlooked in the explainability discussions. There's a difference between global explainability and local explainability. So global explainability is roughly what are the rules of the game here? What are the rules that transform inputs to outputs? What is it that makes it the case that when this data is given, these kinds of outputs are offered? That's global explainability local explainability is why

did this particular input lead to this particular output? Why did the medical data of this particular patient lead to this particular medical diagnosis? Now, both are really important. One thing that's really important about the global explainability stuff is that if you don't know what the rules of the game are, that transform inputs to outputs, then you can't ask the following really important question, are the rules of the game, just, fair, good and reasonable? If the rules of the game are a black box, because it's too mathematically complex for us to assess, we can't ask that really crucial, ethical question.

Jennifer Cohen 12:58
Okay. The title of your book, again, is ethical machines, your concise guide to totally unbiased, transparent, and respectful AI. How can AI become totally unbiased?

Reid Blackman 13:10
It can't that was sort of tongue in cheek, it can't become totally unbiased, there's always going to be differential impacts. Discrimination comes in degrees, it's not an on or off switch. So that title is mostly tongue in cheek, and I can tell you the sort of the story about how the title came to be, which is not totally up to me. But it can't be you're going to have some degree of bias, the issue is not whether you can eliminate the bias, the issue is the extent to which you can mitigate it. And ideally, you're involving AI in a way that mitigates the bias relative to the best human judges.

Jennifer Cohen 13:46
I want to talk about some of the risks that you've already brought up, you brought up racial bias and privacy. Are there other risks to unethical AI?

Reid Blackman 13:58
Yeah, look, I mean, the ethical risks are as varied as the use cases. So to take a non-healthcare example, again, if you're developing a self-driving car, you're going to be using AI. And there the main ethical concerns are not discrimination, privacy violations or black box algorithms. The main concern there is killing and maiming pedestrians. I talked about privacy in the context of the training data and in the kinds of inferences that you might make about people using machine learning or AI, just using those interchangeably. But you also might use it for facial recognition technology for surveillance purposes. And now you've got privacy violations of a different kind. That's the result of the particular application that you're engaging in. So there's plenty of ethical risks, this countless ethical risks and different kinds of use cases are going to give rise to different kinds of ethical risks to varying degrees, which suggests that organizations need a way of assessing comprehensively the ethical risks for a particular use case for an AI that they want to use.

Jennifer Cohen 14:58
Okay, let's get to that. So At the end of your crash course, you identify a number of areas for companies to focus on when developing a kind of risk program. And I want to ask you about two of them. One is defining your AI ethical standards. And one is ensuring compliance with your program is supported by appropriate financial disincentives. So how do you help companies concretize ethical standards?

Reid Blackman 15:24
One thing that a lot of companies are inclined to do, but I think this is a good first step. It's not absolutely necessary. But I think it's a very good first step is to articulate the ethical principles that you want to guide your design, development, procurement, and ultimately deployment of AI. Now, one thing that organizations tend to do is they offer extremely high-level principles that they can't put into practice they can't operationalize. So for

example, they'll say things like, we're for fairness, and then they come to me and they say, okay, how do we operationalize this? What do we do with this, and part of the problem is that that principle is at too high of a level to be put into practice, to bring out that it's too high level, just consider the fact that the KKK, agrees with that principle. At that level, right? They'll say you give them a piece of paper that says the word for fairness, and they'll sign it, they have a phenomenally different conception of what fairness consists in. But at that high level, yeah, they're under that umbrella. Which means that if you're going to articulate what your principles are in any kind of meaningful way, you better be able to exclude certain kinds of people from your articulation of that principle. The way that I do that with clients is to say, you're not allowed to articulate a principle without specifying the things that are that are just off the table. Now, by virtue you're endorsing that principle. So to put that in a sort of an abstract structure, it looks like this, because we value x, we will never y and we will always z now to give some color to that, because we value your privacy, we will never sell your data to a third party. Right. And if you do that with each, you know, you take a principle and you push clients to say, no, no, you've got to tell me the things that are just off the table. Now the principles have to specify at least guardrails of action. And if you can't specify any guardrails, then I don't believe that's a principle you actually uphold or strive for. So that's the first thing, the main thing that we focus on when we talk about articulating ethical standards. And then there's, there's ways of getting at what are those things that are off the table? So for instance, I'll ask clients what their ethical nightmares are, as opposed to their ethical goals? Because we're trying to specify guardrails or do the things we never want to happen. So it's not about what's the rosy ideal look like? It's what does ethical nightmare look like? Okay, if we know that's the ethical nightmare, those are the ethical nightmares. Let's start with thinking about the things that we will always do or that we'll never do to massively decrease the probability of those nightmare situations happening. That's one bucket. The other bucket is more complex, which is something like developing what I call ethical case law. So invariably, what happens in discussions with senior leaders when we're developing these principles is that there are some principles or some cases where there's just disagreement among the team. They're not sure where they stand on it. Will we never sell that to a third party. Well, what about these conditions? What if sharing that data could help many lives, millions of lives what you know, we've got health care data related to COVID. And if we share it, we could help the world by sharing that data, because that data is relevant to say, developing a vaccine or something along those lines. What you get then, is a fodder for a different kind of discussion, which is a discussion about more complex cases. And then working with that team to come to a conclusion about where they stand on those cases. And those can serve as further points of guidance when they come across new novel cases at their doorstep. So that's developing ethical case law.

Jennifer Cohen 19:00
Okay. And on the second area, what sort of financial disincentives are you envisioning?

Reid Blackman 19:06
So look, one of the problems is that there's a number of people in the space, or who first come to the space and they think we need tools for this, they usually mean something like tools used by data scientists to ferret out bias or to get post hoc explanations that have otherwise black box models, or to sufficiently anonymize the data in order to respect people's privacy, things like that. Now, there's one issue about whether those quantitative tools are sufficient to the task at hand, they're not. But there's another about whether the data scientists and developers and product owners etc, are sufficiently incentivized to actually use them. They've got a lot on their plate, they're trying to do a lot of things, and what we're talking about as avoiding ethical pitfalls, but their main goal is not to avoid ethical pitfalls. Their main goal is to make a really good AI a really good machine learning model that achieves or solves the problem that they're to trying to solve if their quarterly evaluations or their annual evaluations that relate to their being promoted, or bonused doesn't take into account at all how well they've dispatched any duties related to using certain kinds of tools to help them, identify bias, mitigate bias, and so on

and so forth, then they're not going to use them. They just won't take them seriously. Some people, I mean, there's a handful of people who will, but unless it's built in to how an organization thinks about, you know, the roles and responsibilities and how those are rewarded or punished when people fail to live up to those responsibilities, you're not going to get serious adoption of AI ethical risk, procedures, tools, techniques, and so on.

Jennifer Cohen 20:46
Okay, so Facebook has put out what they call responsible innovation principles, four of them, never surprise people, provide controls that matter, consider everyone and put people first. And I also saw the Department of Defense had developed some principles for AI research, responsible, equitable, traceable, reliable, governable. What do you make of those is that, does that move in the right direction? Are those as you say, too high levels?

Reid Blackman 21:17
That's just buzzwords. You know, it's the kind of that anyone can say, If anyone can say that, you know, it's a problem. Right? Equitable, maybe I mean, look, put people first. Okay, which people? When? How first should they be first relative to what? What I think about Facebook, they have the same sort of thing with regards to their mission, which is to connect people, okay, which people do want to connect white supremacists, fascists, people who want to engage in genocide in Myanmar, and what's you know, which people do you want to connect? And are there any people that you don't want to connect? Which people do you want to put first, what does it mean to put them first, if you can't dig deeper and specify what that means, and it's just it's meaningless, it's way too open to interpretation? And different people with different motivations will come to us principles and argue that their use case or their particular use for it satisfies the principles and they will be able to find some way or another of saying how this is fair or puts people first, etc. governable? Okay, by whom? Who is it governed? By? What are the standards of governance? What are the methods of enforcement for governance, if you can't fill any of that out, then just saying, well, for governance doesn't mean anything.

Jennifer Cohen 22:23
So interesting. I really appreciated the way you laid this out in your AI ethics Crash Course as two decision tree branches. One is ethics, not for bad. Yeah, as opposed to ethics for good. But I was also intrigued by your title, which I understand is not always up to the author that AI can be respectful. And so I'd love to get your thoughts about incorporating more human types of values, ones that promote the goal of your former field of philosophy of human flourishing, like empathy and stuff is that something that is coming where AI can be respectful, and empathetic and have other more sort of positive values.

Reid Blackman 23:06
So the title of the book is ethical machines. That wasn't my number one title, but it was my number two title. In terms of my preferences, the reason that it's called ethical machines is not because I think that machines can at least now and probably never will embody certain kinds of psychological attitude like respect. But because there are ethical decisions that are embedded as it were in the machines, such that the way the machines operate, can have an ethical impact independently of the intentions of the user of the AI. Okay, so that's a bit abstract. So let me try to put some color on it. It looks like let's say, for the sake of argument that a screwdriver is neither good nor bad. It doesn't have an ethical character, if you're going to use it to build homes for the homeless, great if you're gonna use it to build concentration camps, not so great. So if it has an ethical character, it inherits that ethical character from the intentions of the person who's using the screwdriver. So the tool if you like, is ethically neutral. And the sort of ethical impact all depends upon the intention and goals of the user of the screwdriver. Okay, put that to one side. Is that the case with AI or machine learning? No. So let's go back to the bias case. The data scientists that hopefully more than did that as scientists will make certain kinds of substantive ethical decisions

about who the AI is going to impact and how it will impact them. They'll make certain decisions about whether the model is biased or not. They'll have made certain decisions about what the appropriate strategies and tactics are for bias mitigation. They'll make certain decisions about whether the thing is sufficiently debiased such that it's ready for primetime, or they just won't make those decisions at all, which is still a way of embedding a decision embedding an oversight into the model into the AI. Now you take that model, and you sell it to someone in HR. And now HR is using it to hire people or to vet resumes. And it turns out that the model is discriminatory against, say, Black women. But that's not because of the intentions of that person in HR. It's because certain kinds of ethical decisions were made that affect the kind of thing that the tool is such that the person wielding it will necessarily create certain kinds of discriminatory impacts independently of their intentions to do otherwise. That's at least one way in which I think AI models are interesting, to put it mildly, from an ethical perspective, because they're not ethically neutral in the way that screwdrivers are, which arguably, don't have ethical decisions embedded in their design and development.

Jennifer Cohen 25:46

Right. Let's turn to lessons from health care's ethics risk management history and their applicability to AI ethics. So in a recent article of yours in Harvard Business Review, you discuss how more companies are recognizing the need to address AI ethics issues, and you write a solution is hiding in plain sight. Other industries have already found ways to deal with complex ethical quandaries quickly, effectively and in a way that can be easily replicated. Instead of trying to reinvent this process, companies need to adopt and customize one of health care's greatest inventions, the institutional review board or IRB. So IRBs started in the 70s. And this was on the heels of horrifying revelations about extremely unethical treatment of people. Really, under the guise of medical research, the most infamous being the Tuskegee syphilis study where men were lied to about the so called study they were in they were lied to about available treatments that could have helped their condition. So the IRB was put in place by law to oversee research proposals, to make sure certain protections were in place, that people were fully informed that they had consented, and that the design was put together to try and cause the least amount of harm feasible. And now IRBs are in every university and hospital that conducts research. Can you discuss how the IRB model can be adapted to AI ethics?

Reid Blackman 27:10

Yeah, so there's a way in which I don't mean to say that it's immediately applicable, and we just could just use the existing IRBs as they are and just apply them. But the main point is that what the field have recognized was that there are really important qualitative ethical decisions that can't be simply left in the hands of the researchers, and that a certain kind of panel of experts needs to be assembled to ensure that certain standards are met. And that at a high level is what's going on. The tech community has not realized this yet. They think, oh, fairness, no problem. There's a bunch of technical tools we can develop to solve for fairness. And that strikes me as utterly absurd on its face to think that people with MDS, PhDs in the medical sciences can't sufficiently think of the ethical considerations in AI research and solve for them, but that data scientists and engineers can, that's insane to me. So, rather than saying, Well, look, we just need different technical tools. Let's look at healthcare. And what did they do in the wake of ethically horrific, or to put more mildly ethically problematic cases, they created IRBs to play a certain kind of oversight function. And that's exactly the kind of thing that we need for AI. Now, I know that there are lots of problems with IRBs. But I think, I think everyone agrees better to have an IRB than to have nothing and let anyone do what they want. So same thing with the development of AI, where people are doing a lot of research, and they are letting these AI algorithms fly, and then sort of, we hope, we just hope that nothing bad happens. That's insane. We know what happens. Let's see what this algorithm does on the newsfeed on Facebook and see if it causes depression among teenage girls. Oh, it does, our bad. You know, there's, that's crazy. That is a kind of experimentation on people. Yeah. And it needs a kind of oversight that's akin to the IRB.

Jennifer Cohen 29:12
Yeah. So a key aspect of an IRB, which you discuss is membership, you know, who should be on them? Should it be a group of people internally or should it be an external IRB that's hired and you have a recommendation that like medical research IRB, someone unaffiliated with the organization should sit on the IRB, which I thought was fascinating. How does that go over? And what type of person are you envisioning?

Reid Blackman 29:38
Yeah, it's tough. I mean, there's, you know, what I would like if I had a magic wand, what would occur if I would wave that magic wand? And if that would happen, then I would say, you know, external IRBs would be the ideal. I don't see that happening anytime soon. If ever. Companies are, depends on the size of the company, and different companies have different kinds of cultures, but some companies are open to having a member from outside their company sit on their ethics board on their AI ethics board. And as for who it ought to be, well, it's probably in a very, I mean, in some cases is going to make sense to have, say, a civil rights attorney, given the kind of product that they create. But in some cases, it might make sense to have a medical ethicist, you know, let's say they're doing AI and healthcare AI. And you know, they create AI's that diagnose illness, it might make a lot of sense to have a medical ethicist involved, might make sense to have a sort of general ethicist, say someone with a PhD in philosophy, who works on ethics, it's going to vary who the appropriate person is. But I think it's fairly crucial to have someone who has an expertise and ethics. Usually they're going to be outside, they don't have to. But the problem is, there's going to be this as sort of call back to the issue about ethical risks as a result of use cases. And those ethical risks being as varied as the use case, there's tons of ethical issues that pop up, that, frankly, most people don't know how to think through. And they're not to blame, they just haven't been trained to think through those kinds of things. And people with a certain kind of training and ethics know how to navigate that space to help people engage in those discussions in a fruitful manner not to sort of act as a kind of priest from upon high insight, this was the right thing to do, but to get people to see the issues clearer than they would otherwise so they can make a more informed decision.

Jennifer Cohen 31:19
So you in addition to your philosophy background, I know you've taught medical ethics, do you see the ethical frameworks that developed around medicine, the four principles for clinical medicine, for medical research, there's a slightly different ethical framework where you're, you're testing a hypothesis, you're not trying to ameliorate a particular individual's disease. Do you think that those different health care ethics frameworks are applicable to working through an AI ethics problem?

Reid Blackman 31:53
That's a good question. I haven't thought too much about that. My best guess is yes, there's going to be different, you know, I'm not, I have trouble wrapping my head around frameworks. It's just not how I think, as a trained philosopher, when I think about what progress looks like in medical ethics, as a philosopher, I think about well, look, what we do is we're really engaged in analogical reasoning. So for instance, this case of letting someone die is perfectly permissible, this other case, is sufficiently similar. So we should think that it's ethically permissible in this case, as well. And you do that sort of thing. What's the difference between someone refusing treatment and someone being assisted to kill themselves? And thinking through the relevant distinctions, there is how we get progress, not by applying some framework. Honestly, I have trouble wrapping my head around frameworks for ethics. That said, I get that there has to be something especially when you're not doing sort of highly abstract meant for 1 percent of the 1 percent of the population who works on this kind of thing. Probably less than that, actually. So yeah, different kinds of frameworks going to be appropriate for different kinds of use cases.

Jennifer Cohen 32:58
Allright let me ask you about a specific healthcare AI ethics, case study Optum health, that became the subject of an investigation when an investigation determined that one of their algorithms was racially biased. How much regulation is go, I think the regulating bodies in that case, were New York's financial services department, because I guess Optum health They're a provider, but they're owned by an insurance company. So maybe that kicked in that department in terms of regulation. But is there AI regulation at the federal level? Is that being proposed? Is that coming in the same way that it's arrived in Europe?

Reid Blackman 33:35
Well, AI regulation hasn't quite arrived in Europe. Europe has GDPR, which is data protection regulation. So it's more data specific than it is something like machine learning models specific. Okay. There are AI regulations coming down the pipeline in the EU. There is some arguably, there's stuff about AI models in the GDPR. But it hasn't been actually tried in court. So nobody really knows. Nobody really knows what it means or if it's applicable, or what will fly and what won't. So there's more by way of regulation coming out of the EU by far than the states. There's also the Digital Services Act or Digital Marketing Act, it's coming out in about I think, about a year and a half is gonna start to get enforced, which is AI regulations around things like discrimination and explainability. The US we've got basically nothing. There's the CCPA California Consumer Protection Agents Protection Act, or is it Privacy Act, which is akin to the GDPR. But if you're not in California doesn't affect you. So most Americans don't live in California, it doesn't affect them. There's sort of pockets of regulation, this state or this city, outlawing facial recognition software, for instance, there's talk by some government agencies that they're going to start prosecuting or investigating alleged discrimination when people use AI so that's using existing regulations against discrimination that could come to bear. We haven't seen too much enforcement there in terms of a coherent, robust Federal AI regulation for the US, it's nowhere in sight, which is not always scary by itself. But the thing that I've started to talk more and more about is, first of all, we're at the beginning of AI, it's gonna get developed, and it's going to, adoption is going to increase, and it's just going to become more and more of a mess. And then you're going to throw things in there, like quantum computing, which I know, you know, we don't need to get into, but quantum computing and blockchain and all these technologies are gonna get thrown in, and they're gonna get combined with each other. So web three is an attempt to combine blockchain technologies with AI, for instance, or at least lots of applications are. So in quantum computing, you want to use quantum computing and AI for precision medicine. And forget about having regulations or on any of the stuff anytime soon. So it's fairly scary, in my opinion.

Jennifer Cohen 35:49
Yeah. Okay, let's turn then now to data privacy. You say in your AI ethics, Crash Course, the fuel of AI is people's privacy. And you talk about the main ingredients to ethical privacy that people should be thinking about. The first three are, how transparent you are about what data you're collecting, and what you're doing with the data, how much control people have over their data, and whether people have consented to give their data? Can you discuss how well you feel businesses are doing addressing those concerns now?

Reid Blackman 36:23
Oh, yeah, they're I mean, they're fairly awful. They are potentially good, their ordinances are good at taking about privacy from some perspectives, but not others. So one thing that is a bit frustrating when talking about privacy is that there's various ways that people address the privacy issue, some just, when they're talking about privacy, they're just talking about regulatory compliance with the GDPR or CCPA. And they think privacy is sufficiently respected on the condition that they're compliant with regulations. Now, that can't be enough. Like I mentioned,

for all those people who live in the US, and not in California, they have very little protections with regards to their data. Then there is this goes beyond what I said in the crash course. But there's a sort of a passive conception of privacy and a active conception. So when you think of the passive conception, think of it as something like people's privacy is respected, you know, your privacy is respected on the condition that someone else secures that privacy. How would they do that? Well, if there's data about you, then from a cybersecurity perspective, your privacy is respected on the condition that only people who want to have access to that data do in fact, have access. And those people who ought not to have access do not have access. So that's just your privacy is respected on the condition that the cybersecurity people do their job in controlling access to your data. Okay. Another way of thinking about privacy, also passive is that your privacy is respected on the condition that your data about you is anonymized with a sufficiently low probability of being deanonymized. So, again, very passive, you don't anonymize and have your data, they whoever they are, they anonymize your data. Now, when it comes to those two buckets, the cybersecurity perspective of access control and the anonymization stuff, which is mostly done by data scientists, your passive with regards to your data, right. This is contrasted with what I would call an act of conception of privacy, which is more along the lines of illegal conception of privacy where a right to privacy is not the state of being a passive state of being as an active state, it's a capacity that you can exercise, you can exercise your right to privacy. So to take a non data example, you can choose to close the shades in your bedroom or not, that's a choice that you make as a capacity that you have to draw the shades or not. If someone knocks on your bedroom door, you have the capacity to let them in or not to let them and that's you exercising your right to privacy. So it's exercising a certain level of control about who has access to you and your bedroom. So the analog in the data case, could be privacy with respect to your data would be your ability to control who has access to your data under what conditions for how long what they can do with it, and so on. I would think that's something like the ethically maximal conception of data privacy, it's when you've got control of your data. Another sort of analogy here is, when you think about privacy and autonomy in the medical sphere, you might think about things like bodily self determination, you have control of what happens in into your body. Certain kinds of threats happened to that lately, as we both know, but that's what a lot of privacy and autonomy looks like as bodily self-determination. And then the analog and this is actually a phrase that's used in Germany, not here because it's too intellectual. But there's a thing called informational self-determination, where you have control over what happens to information that is to say data about you. How well are people or how well are organizations doing with regards to that poorly, phenomenally poorly

Jennifer Cohen 39:57
And underlying the legal concept of privacy, our property rights? Yeah. And then the examples, you're giving you the idea that you own your body, you own your bedroom. Yeah. Should people own their personal data?

Reid Blackman 40:11
So there's the standard line. And then there's the controversial line, the standard line in the AI ethics community, and admittedly, the one that I espouse in my book, because that book is not the place to go into various philosophical subtleties. Is that Yes, people, you know, it's my data, I should have control over my data. That's my private data. That's my private information about me, I should have control over it. To the extent that you take that data without my consent, you violated my privacy. That's by far the standard line in the AI ethics community. I don't think it's crazy. Be a proper philosopher here. And I'll say it's a not unreasonable view. My own inclination is to think that there's a very quick move here from data about me to my data, that there's an inference. This is that about me, thus, it's my data. And that just strikes me as a straightforwardly bad inference. So for instance, if you note that, you know, you're sitting in the cafe, and at 10 o'clock, you see me walk into the cafe and you write down in your notebook Reid entered the cafe at 10am. That's data about me. But I can't go over there and rip the page out of their notebook because that's my data that wouldn't fly so and then the AI ethics community, I think that

distinction is missed, where they think, oh, it's, it's data about me, it's my data, you're taking my data. So I'm disinclined to think that really, respect for privacy requires us having ownership over our data. But that's an unpopular view.

Jennifer Cohen 41:35
That's fascinating. I'd love to just pick your brain on this observation that there's a recent Brookings article, entitled Why is AI adoption in healthcare lagging and it goes through first Eric Topol's book deep medicine, which was an incredibly optimistic look at all the different ways AI was going to help patients and even improve the lives of doctors. And in this article, it has this incredible chart of 20 industries, showing which percentage of their job postings required AI skills and healthcare was second from the bottom, there were more job postings requiring AI skills in agriculture and forestry, in in the healthcare sector. Does that align with what you're seeing in health care?

Reid Blackman 42:19
Yes, I mean, I have better access or better insight into how much healthcare is taking AI ethics seriously, as opposed to AI generally. And they don't, and perhaps part of explanation is that they're just not doing AI yet. And perhaps part of the explanation for that is that they don't know how to control it. A few years back, I thought, You know what, I think that healthcare is going to be the place that adopts AI ethical risk programs first, because healthcare speaks the language of ethics, right? There's IRB, these hospitals have ethics committees, there's ethical codes of conduct for healthcare practitioners. They're not scared of the word ethics in the way that frankly, lots of other industries are scared of that word. They don't like to say the word ethics, the same responsible, maybe trustworthy, but not ethics. But I was wrong, they are not. They're not doing any of it. And part of is that they don't have the infrastructure to deal with this sort of thing. The risks are phenomenally high, obviously, how could they be higher, and they don't have the proper infrastructure for oversight. They don't have the right people who are trained in identifying ethical risks of AI. And so I think that, you know, there's just a bunch of legacy systems in place, and they're relatively conservatives. And since they know, they're aware of the ethical risks of AI, I think that that's one of the things holding them back from really adopting AI.

Jennifer Cohen 43:37
Okay, so my last two questions, I want to look to the future. You've brought up the idea of training and lack of training in a number of contexts. Can any of these thorny issues around AI ethics be addressed? In your old stomping grounds the academy, before people enter the business world, I was reading about Norbert Wiener, who while at MIT was arguing for a new type of computer programming ethics back in the 1940s. My question is, is there any movement in the academy to provide more or any ethics training for computer science?

Reid Blackman 44:12
I think the answer is yes. I mean, I haven't seen a tremendous amount of it. And what I have seen is mostly at the, you know, the most elite places, MIT, Carnegie Mellon, where they want to introduce AI ethics or data ethics or something like ethics into the curriculum. I am not a huge proponent of this, I'm not against it, sure, go and do it. But the idea that you're going to give undergraduate, you know, majors in physics or computer science, a sufficiently robust understanding of ethics such that they can then actually apply it in their work. That just strikes me as relatively preposterous. Again, it's like, oh, yeah, the healthcare world. They need IRB, they need real oversight by a committee blah, blah, blah. But data scientists just need to an undergraduate course in ethics, that'll do it that strikes me as foolish. I think that they've got a certain expertise in computer science and data science and engineering and coding. And one of the answers the ethical problems is not let's give them an ethics course, it's let's get cross functional expertise involved solve the problem. So for instance, let's get ethicists involved or civil rights attorneys or medical ethicists and so on and so on. You know, I'm all for giving them a course on ethics so

that they can at least learn ideally, that, frankly, ethics isn't bullshit, which is what they think most of the techies, if it's mathematical, if it's quantitative, it deserves our intellectual respect, if it's ethics, and it's squishy, and subjective, and just opinions and his emotions. And it's not, you know, the real province of intellectual inquiry, its mere opinion, its taste its sentiment. And if I didn't want an ethics course, to accomplish anything, for those budding data scientists and engineers, it wouldn't be for them to be able to do ethics well as a word it would be to get them to see that it's an area in which there's real expertise, and having those experts involved in their decision-making processes is necessary and fruitful. That should be the goal, not training them to be ethicists.

Jennifer Cohen 46:14
Okay. Okay, my last question, there's a ton of anxiety about AI, there has been from the beginning that it's a technology that's going to replace human productivity, or one that amplifies our worst traits. Are you optimistic about the future of AI? And where do you predict AI will go in the near term?

Reid Blackman 46:35
I don't know. Am I optimistic? If I had to put my money on something, I'd say there's going to be a lot more wrongdoing before there's not, you know, we're nowhere near federal regulations. We're inching towards very small state why regulations like CCPA, New York seems to be doing something, especially on hiring. So you know, we're gonna see progress here and there I think the most optimistic one can be is that the EU is going to pass regulations in the near future, and seriously enforce them. And that's going to force multinationals, American based companies say that operate in Europe, they're gonna have to comply with those regulations, they'll find it easier to comply with them across borders than just limiting it because it's just too hard to slice it all up like that. That's probably the best hope we have. There has been over the past several years, a massive increase in companies taking the issue seriously. But a massive increase when you're starting from something very, very tiny, still leaves lots of risk on the table still leaves lots of companies doing lots of risky, ethically risky things. So I think we'll see an increase. I think we're gonna see a lot more bad stuff before we see safety.

Jennifer Cohen 47:46
Reid Blackman, thank you so much for a fascinating discussion and for your critical work and improving this emerging field that's already a part of all of our lives. Best of luck in the future.

Reid Blackman 47:57
Thanks so much. It's nice talking to you.