***The Good Bot: Artificial Intelligence, Health Care, and the Law —***
**AI Discrimination and Emerging Best Practices – Part 1**
**Hosts: Brett Mason and Jim Koenig**
**Guests: Alison Ground and Chris Willis**
**Recorded: 11/07/23**

**Brett Mason:**

Welcome to *The Good Bot*, a podcast focusing on the intersection of artificial intelligence, health care, and the law. I'm Brett Mason, your host. As a trial lawyer at Troutman Pepper, my primary focus is on litigating and trying cases for life sciences and healthcare companies. However, as a self-proclaimed tech enthusiast, I'm also deeply fascinated by the role of technology in advancing the healthcare industry. Our mission with this podcast is to equip you with a comprehensive understanding of artificial intelligence technology, its current and potential future applications in health care, and the legal implications of integrating this technology into the healthcare sector.

I'm excited about today's episode, which is part one of a two-part series focusing on algorithm fairness and bias and discrimination when using AI. The first part of the series highlights a conversation between three of my colleagues from Troutman Pepper. Jim Koenig, the global co-lead of privacy and cyber practice here at Troutman Pepper. He also sits on our AI task force. He's joined by my colleague Alison Grounds, who is also a partner at Troutman Pepper. She's also the chair of the innovation committee and one of the leaders of the generative AI task force here at Troutman. And Chris Willis, another one of our colleagues who's the lead of the Consumer Financial Services Group. I hope you enjoy the conversation between the three of them talking about bias and discrimination in AI and emerging best practices to address that bias and discrimination.

**Jim Koenig:**

Well, welcome, everybody. I want to welcome you to the next in the Troutman Pepper series, Managing AI: Risk, Reward, and Regulatory. Today's session will focus on AI discrimination and emerging best practices to avoid bias and discrimination.

So now let's focus, at least for today, on discrimination and bias. I've set the stage to talk a little bit about AI, the uses, industry-specific, some of the benefits, and what's happening in the law. But for those of us who are not everyday experts in dealing with it, okay, I didn't spend a lot of time in it, but to be able to know it when you see it, we're going to turn to that next.

We're going to talk a little bit about the technical and procedural ways that bias be creeped into the data set and into the training of the algorithm or the output as well. It's not always obvious. Alison, why don't you talk us through just some examples that come across industry that'll help everyone be on a common page about understanding a little bit better what we're talking about when we talk about data set, algorithm, or output bias.

**Alison Grounds:**

Sure, Jim. The theme here across all these different industries and use cases is that the historical bad data in, bad data out, right? So you're really looking at what's being analyzed for the AI to help make a decision or a recommendation. So you can have certainly historical hiring data that may be a little bit skewed. A great example I give all the time is if you were to ask the AI based on historical data to tell it and predict who would be a good partner at the law firm, my face in my bio would not appear. So you're really looking at the data set that's being used to train the algorithm, and you can see that across industries. Medical data sets could skew towards one ethnicity or socioeconomic class, depending on the access and accuracy and bias in that data set.

In the financial space, which Chris can speak to more in a more detailed way than I can, certainly potentially looking at past loan approvals may skew the data in ways that may favor certain neighborhoods or demographics or discriminate against or show biases against others. You've seen this in the criminal justice system as well, and there's a specific reference to this in the executive order, looking at past arrest data and other information regarding our criminal justice system could incorrectly input some bias or discrimination and sentencing recommendations. And there was some publicity about this not too long ago. This is all kind of classic AI use cases that have been around for a while, nothing really new. But as, Jim, you mentioned, really getting more attention now that you're seeing anyone with access to a computer able to play with ChatGPT and use generative AI and search engines and other tools that are out there now.

But another one that I think I've seen and kind of moving it over to the generative AI space, if you took my earlier example of asking for it to help me find and recruit a future successful law firm partner, if you also asked it to generate an image of what that partner would look like, that image would likely be white and would likely be male. So that's based on this historical data that it could be using, especially when we talk about maybe some unique issues of bias and discrimination that can creep into the generative AI data set, which is, it's trained on things we're not exactly sure about, but potentially vast amounts of information from online resources that could be limited in the diversity and inclusion and the data sets that it has available to it.

So I talked a little bit about some of the historical information leading to bias and discrimination. Also, one of the possibilities and one of the issues here is that the data, if it's skewed too much to only recent data, 90% of data being created in the last two years can also skew those results and may generate an information or bias that's not necessarily accurate. Additionally, unlike some of the traditional AI models where it may be easier to pinpoint the particular bad data or where there may be over-reliance on an inappropriate criteria, the complexity of the large language models and neural networks used in generative AI versus more traditional machine learning makes it even more difficult to look behind the curtain, if you will, and see what's causing that bias or discrimination.

These issues require some solutions and some thoughtful strategies to make sure we are addressing bias and discrimination. So certainly, the executive order contemplates some pre-release testing of certain large language models and generative AI tools to make sure that they function as intended. Other suggestions that you see from the other frameworks and documentation around this space is, of course, continual auditing to make sure while it may be

working as intended initially, that it doesn't start to skew and move in a different direction as it continues to be used and incorporate additional data and learning.

Certainly, user education and awareness, this is a big thing we've done here at the firm when we rolled out our own generative AI tool. Letting people know what it was good for and not good for, letting them be aware of their ethical duties and obligations. And then, of course, including some attorney oversight for the tool that we are using internally and other tools we may use. Making sure you have the appropriate alignment with the human intentions, and that would be true across industries as well. And then considering the option of being model agnostic and being open to testing different models that may be better suited in combinations of models working together to help potentially reduce bias and discrimination in the use of AI.

**Jim Koenig:**

Now that we framed discrimination and bias, how it can create them and understand a little bit better in different industries. So at least hopefully each of you can relate in one way or another to understand the problem just a little bit clearer and a little bit better. Now let's focus on the evolving best practices to prevent discrimination and bias. With that, let me turn to Chris to talk a little bit more about this area based on his superpowers and experience. Chris?

**Chris Willis:**

Yeah. Thanks a lot Jim.

As you've heard Jim say, we have a situation here where there's a lot of skepticism of AI by the media and by politicians, and therefore also by regulators. You can scarcely open any news site without seeing some kind of story about something bad that AI did or can do. And many of those focus on the possibility of discrimination. You've seen news about it in terms of facial recognition, algorithms, and credit scoring models and fraud detection models and things like that. So we have to understand that we're operating in this environment of extreme skepticism for regulators.

In my area, which is financial services, one from the current chair of the Federal Trade Commission, which of course the FTC sees a large swath of the American economy, who commented earlier this year that the use of artificial intelligence can "automate discrimination" and make it obviously much worse that it otherwise would be without AI was the subtext of the statement. And we have the Consumer Financial Protection Bureau, which is the agency that I deal with on a regular basis in the consumer finance area, who loves to use the catchphrase black box discriminatory algorithms to describe AI and machine learning models that are used by the financial services industry. So when your regulators are talking like that, you know you sort of start off with a presumption against the use of these technologies. On the other hand, there's an incredibly strong use case for them because they are more efficient and more effective and can be more fair than the models that they're replacing, sort of traditional logistic regression models.

So what I wanted to do was talk about the use cases of specific AI models that are used to predict something, things like in the credit industry, whether someone is submitting a fraudulent application or not, or whether a person who's applying for credit is likely to repay the credit or not. Those are very strong use cases that are heavily in use in the financial services industry,

and I think we can learn from those in terms of what are our best practices for adopting AI really in any industry, because the technology really works in a similar way. You're going to hear echoes of some things that you just heard Alison say just a moment ago in my comments.

But let's talk about what the laws are here, what law applies to this. I mean, we can all sort of generally agree that bias and discrimination are bad, but what law applies to that? Well, of course in the credit world, we have a specific anti-discrimination law called the Equal Credit Opportunity Act. There are other specific anti-discrimination laws at the federal level like those dealing with employment, for example, or housing. But in addition, one thing that I want to make everybody aware of is, there are lots of state anti-discrimination laws too, and some of those are specific to particular types of transactions like credit or housing. But there are a number of states that have sort of all-purpose, generalized anti-discrimination statutes that apply to any kind of contractual or other relationship between people or companies.

So one example is the California Unruh Act, which prohibits discrimination on a variety of protected characteristics, including all the ones that you would normally think of and some you probably aren't aware of. Things like race, ethnicity, gender, age, things like that, but also things like immigration status and religion and all kinds of stuff like that. And New York has a human rights law that's very similar, that just essentially prohibits discrimination across the board. So you have those two states with those laws in place and who also happen to have very active state attorneys general in terms of wanting to bring cases dealing with the subject of discrimination.

So I don't want you to think that our legal landscape is composed entirely of federal laws. It certainly has them, but these state laws can play a major role too, and there's an incredible opportunity waiting there for those state attorneys general to use their state anti-discrimination laws in any way that they see fit. So when we think about best practices for how to prevent bias and discrimination from AI models, what we're really trying to do is play to the audience of both the federal and the state regulators who may be looking at our efforts to adopt this technology and what we think they think will find credible in terms of our efforts in that regard. So understanding this sort of environment of heavy scrutiny is the first step to understanding what we need to do.

I've got a series of suggestions for things that a doctors of AI models can put into place that some may be appropriate for some industries and not others, but they all stem from the basic point that you heard Alison make a minute ago, which is that the place that bias and discrimination can come into a machine learning model or an AI algorithm is really from two places. One is from the data that is used to train it. These models get data sets and they train on them to look for correlations between variables or the association of several variables and the outcome that is seeking to be predicted. The outcome of, is this user going to click on an ad and be interested in a product? Or will they repay a credit transaction? Or is this a fraudulent application? Or things like that.

So they look for associations and correlations, and the training data is where they do it. So when we look for sources of bias or discrimination in these models, it is very frequently in what the training data is, and I'll talk about that in a minute. But the other source of it is, what are the attributes that we use? In other words, are we letting the model train on attributes of a particular person or application, or whatever the case may be, that may themselves, even individually, be strongly highly correlated with race or ethnicity? Bias detection, really goes towards finding

where those capabilities for bias or discrimination exist in the training data set or in the attributes that we use.

So for example, if a training data set is small, and in particular, if it has only small numbers of people of protected classes within it, then it will have the tendency to potentially create bias or discrimination in the outcome because of the capability of machine learning models, particularly those used in these kinds of predictions to overfit themselves. Overfitting just means that the model draws a correlation that spurious or idiosyncratic that's not real in the real world, but it appears to be a correlation based on the limited training data available to the model. So if, for example, you have a small number of people from a particular protected class in a development data set, then you can have the model overfit onto characteristics of those people that aren't truly predictive of what you're trying to predict, but nevertheless will appear as correlations in the development data set. So in order to prevent that, the main thing is to have as large and diverse a data set as possible to make sure you don't unintentionally incorporate that bias into the outcome.

The other thing is, in terms of the attributes, certainly in the financial services industry, for as long as I've been practicing, in fact, as long as we've had these discrimination laws specific to credit, there have been certain attributes that the federal regulators have warned against using. A perfect example is ZIP code. So people live in different ZIP codes, and you can even get in a smaller geographic unit than that like census tracts or block groups or things like that. And you would find, if you did an analysis, that there are strong correlations between the ZIP code where somebody lives and their likelihood of repaying a credit obligation, for example. But there's another thing that ZIP code is highly correlated with too, and that's race and ethnicity, because of the segregation of housing patterns in a lot of areas of this country.

So going back many, many years, the federal regulators have warned financial institutions not to use ZIP code as a predictive attribute in credit underwriting models because of the fear that you will basically be underwriting people based on where they live and thereby underwriting them based on their race or ethnicity. Running through that is another sort of undercurrent of thought among the regulators, which is, it's unfair to judge people on an important decision, like will you get credit or will you get housing or something like that, by people who you sort of live near or are associated with.

So for example, let's say you take somebody who lives in a ZIP code that has a lot of people who have poor credit scores, but they themselves are very responsible with credit. Well, if creditors use ZIP codes to underwrite for credit obligations, then the person who lives in that ZIP code who's very responsible with credit never has the opportunity to access credit or has a diminished opportunity to access it just because of the neighborhood into which he or she might've been born or might be living. So you judge somebody not sort of on their own merits, but on the merits of the company they keep or the neighborhood that they live in. And that's thought to be very unfair from a regulatory standpoint.

So I think the bias detection piece of the prevention mechanisms that we have here has to do with not only looking at our development data set, but also looking at the attributes that we're going to allow the model to train on. And are any of those attributes going to be correlated with race, ethnicity, gender, age? Like letting it train on people's birthdays or their names or things like that can all be potential sources of bias. If you let the model train on it, it will find correlations between everything that you give it and the outcome you're trying to predict. So controlling the

information available of the algorithm is one of the prevention mechanisms that we think is important.

When we say diversity and development, there have been a number of instances where model developers were all sort of uniform in terms of their own sort of characteristics. So they inadvertently included attributes in a model that might have the capacity to be correlated with a protected characteristic. The bottom line here is that if there's diversity among the model development team, they're more likely to spot those potential problem areas with sort of new and innovative attributes than sort of a very homogenous team might be able to do. So having that diversity I think is an important risk mitigation as well.

Transparency, is about making sure that we understand how a model works and why it reaches decisions or scoring outputs or whatever it is that it creates as an output. There are some models where it's very difficult to have that sort of transparency. The good thing in credit is, the types of models that are mostly used in credit have had significant advances in transparency, which is really required because in credit, when somebody gets declined because of an underwriting model, we have to give them an adverse action notice that tells them why they were declined. So the model has to be transparent in order to fulfill that regulatory and legal requirement. But even beyond that, there are lots of opportunities for models to become transparent or to be analyzed so that we know why they're making decisions and which features in the model are the most important to its outcome.

Evaluation metrics is obviously very important, and it may surprise you to know that there are pretty well established methods for testing models of all types. And these metrics go back to before machine learning and AI even existed, but they still apply just as easily to current models to see essentially how they come out, what is the score distribution between, say men and women, or older and younger people, or people of different races or ethnicities. So you can actually see if there's bias or discrimination in the model. And then you can see if that bias or discrimination is related to actual real outcomes that are different between those groups, or if it's an unfair bias against members of protected classes.

Because remember, when we deal with disparate impact, which is the primary legal theory that is involved in this, a business justification is a defense to that. So there are firms that specialize in testing models, and we've had clients develop this capability in-house too by looking at essentially the score distribution, the scores between protected classes, to know whether you have a discrimination or bias problem. So those metrics exist, those testing methodologies exist, and they can easily be transported to other scenarios besides just financial services.

SDLC AI impact assessments is really speaking to a greater emphasis on looking at the impact assessment of an AI model. So in credit, obviously we know what the impact of a model is because it approves or declines people and sets their interest rates, but in other use cases, we have to understand, when I incorporate this algorithm into my operations, what does it do in terms of its impact on my customers or on other external parties? So assessing that impact is a necessary step to understanding how important it is to try to trim discrimination out of a model and how it will impact people if we refuse to do so.

Regular reviews, this is something that we're used to in the credit world because the models that we use in credit have to be updated every once in a while to take into account new macroeconomic conditions or changes in the applicant pool for a particular product. And we can

take that as a best practice for other industries as well. You don't just set an AI model and put it into production and then leave it for 10 or 20 years. You would want to regularly review what it's doing. Is the output as expected? Is it working as expected? And have the environmental factors change such that we need to review the model and make sure that it's still working like it is supposed to? So that's what regular review speaks to. And the frequency of those reviews really depends on the model and the use case, but it's certainly a necessity for any model, any use of AI to regularly review whether it's working as intended or not.

Ethical guidelines speaks to the use of information by the model. What ethical guidelines will we have even beyond our legal requirements to say what data will we use in the model, what notice will consumers have that their data is being used, and what opportunity do they have to find out information about how the model is impacting them? Those are obviously different in different use cases and different industries.

External audits, I mentioned this a moment ago, that in terms of assessing whether there is bias or discrimination in models, there's a role for lawyers to play in terms of looking at the model development process, the attributes that are used, and interpreting the outcomes of these evaluation metrics that I talked about. And there's a role for specialists like statisticians and economists to play in terms of actually conducting statistical reviews of model outputs and their business justifications, if that's appropriate to a particular use case. So there's lots of specialist firms out there that do that, that we work with on a regular basis. And of course, we do our own external audits and reviews in connection with those statisticians, or sometimes by ourselves if we don't have to do the statistical analysis to do it. So there is help out there to design a method of testing and evaluating models. So if you need it, don't hesitate to call upon it.

Education and training I think is really critical, and it really, in my world, has to do with educating and training the people who are responsible for model development to help them understand both the legal and regulatory and the PR and ethical considerations that the company wants to attach to model development, understanding what is the public dialogue about bias and discrimination models, where does it come from and what do we do as modelers to try to avoid introducing bias or discrimination models, and how we would detect it and remedy it if we did. That's what the education and training is that we're referring to there.

Finally, redress mechanisms, this isn't really unique to AI or machine learning. It really has existed as long as there have been litigation and regulatory outcomes finding bias and discrimination in the use of models, even free AI models. So you'll see this in lots of regulatory consent orders, that the defendant in those consent orders is required to pay restitution or redress to people who are negatively impacted by a model that a regulator thought was discriminatory. So a lot of times you'll see financial institutions in particular, when they do their own internal testing or do it with the help of external parties, they will potentially pay restitution to people if they find that there's been a violation of law for which restitution is appropriate. So that's something to consider as we build out our best practices for thinking about how we deal with potential bias in AI.

**Brett Mason:**

Thank you to our listeners. I hope you enjoyed the conversation between my colleagues talking about AI bias and discrimination emerging best practices. Please don't hesitate to reach out to me at brettmason@troutman.com with any questions, comments, or topic suggestions. You can

also subscribe and listen to other Troutman Pepper podcast wherever you listen to podcasts, including Apple, Google, and Spotify. As we mentioned at the beginning of this podcast, this is part one of a two-part series focusing on AI bias and discrimination and emerging best practices. I hope you'll join us for part two of this series, where we will have a fireside chat with industry leader Pedro Pavón from Meta.