**Transcript: Identifying and Addressing Bias in Machine Learning Models on Selection of Candidates from a Policy Perspective**

**Future of Work and Disability webinar originally recorded on November 17, 2020.**

**Watch** [YouTube video](https://www.youtube.com/watch?v=Tq230FBgoiU&feature=youtu.be)**.**

**Vera Roberts:** Welcome to the second Future of Work and Disability webinar in the We Count Digging DEEPer series. I think all of you can see that captions are available today, and we have ASL interpretation who for those who would also like to use it.

I’d like to begin with our land acknowledgment. OCAD University acknowledges the ancestral and traditional territories of the Mississaugas of the Credit, the Haudenosaunee, the Anishinaabe and the Huron-Wendat, who are the original owners and custodians of the land on which we stand and create. I’d also like to acknowledge that you are joining us today from many places, near and far, and acknowledge the traditional caretakers of those lands.

Hello, everyone, I am Vera Roberts. I work at the Inclusive Design Research Centre. I have the honour of moderating today’s session, so welcome again. I am going to tell you about our agenda. So today we will have presentations from our panelists, followed by discussion, and then we’ll have opportunities to have questions from the audience.

Now I would like to take a moment to introduce our guest speakers. We are very fortunate today to have Alexandra Reeve Givens and Julia Stoyanovich with us. Alexandra is the CEO of the Center for Democracy and Technology. It’s a think tank that focuses on protecting democracy, individual rights and the digital age policy. In her spare time, Alexandra serves on the board of the Christopher and Dana Reeve Foundation. She’s also a Mayoral Appointee on the Washington, D.C., Innovation and Technology Inclusion Council.

Julia is an Assistant Professor of Computer Science and Engineering and of Data Science at New York University. She researches responsible data management and analysis. She ensures that fairness, diversity, transparency and data protection in all stages of the data science lifecycle through her research. She is the founding director of the Center for Responsible AI at New York University. This is a comprehensive laboratory that is building a future in which responsible AI will be the only kind accepted by society.

Thank you for joining us today. Alexandra, I am going to let you begin with your presentation.

**Alexandra Reeve Givens:** Wonderful. Hello, everybody. It’s very nice to join you from down here in Washington, D.C. I’m going to share my screen because I have some slides. They were shared ahead of time. For those of you who are not able to see the slides directly, know that the content is going to exactly be reflected in my words, so you can follow along that way if you would like to.

I’m going to talk today about two areas that focus on two particular pieces of this question about bias in the selection of candidates. My focus is going to be on what some of the legal frameworks are, what some potential avenues of argument are against bias, and the types of content that we in my organization are pushing companies to think about when they engage in these issues, and tools that we hope advocates will have.

**[Types of Tools]**

I’m going to start by doing a quick overview of some of the types of tools that we are talking about to help ground the conversation. Some of these may well be very familiar to those of you in the working group. But just in case, I think it’s useful to pause and talk about some of the tools being used and some of the concerns that they raise.

So one of the classic examples of the use of AI and hiring is the use of resumé screening. Tools that help sort fast, stacks of resumés into smaller amounts that may be tailored to a particular job description. The types of concerns that we may be aware of here from a disability and general inclusion perspective is that the tools are often trained to look for traits or characteristics in an existing employee pool. We can think of the risk of that perpetuating existing patterns of inequality in a workforce.

One famous example was a tech company that used one of these tools for their internal hiring processes, and ended up realizing that the samples of people coming through were skewing overwhelmingly towards men. The reason was that the algorithm had been trained on the resumes of successful existing employees, which had really poor gender representation within them.

From a disability perspective, we can think of gaps in a person’s resumé, where they may have gone to a place of higher education, participation on a sports team, you can name it, all the different types of traits that might be inadvertently being factored into the screening here, and how that can end up disadvantaging people whose resumés may differ from those within the established selection pool.

Another example of a tool is the use of video interviews that then report to conduct sentiment analysis or facial recognition analysis on a person’s conduct during that interview.

Another example is the increasing use of games and logic tests in hiring tools. There is a sample appear on the slide for those of you who were able to see it. Here we have, for example, a test for numerical and logical reasoning, which invites the candidate to click on an image which says, which side of these two images has a larger proportion of yellow dots? And what they’re measuring there is the speed and accuracy of your return of that answer.

Obviously, that is a specific example. You can imagine if someone is colour blind, choosing the number of dots of a particular colour on a screen may be troubling. If you think of someone with a mobility impairment, speed of response time being seen as a gauge of your intelligence or your quickness and response is actually not a fair or accurate measure of your abilities.

On the right side of the slide, we see some examples that describe what this tool is reporting to measure; what some of the games in this particular company are reporting to measure. The factors include things like attention, effort, fairness, decision-making, focus, generosity. Some of those too we can be worried about being coded words or measures that will have a significant disadvantage for people who present with a particular disability. I’m happy to talk more about these in the Q&A. We are always working to surface examples of these types of tools and the potential exclusionary effects that they may have, even exclusionary effects that are not remotely intended by the designers, but may well end up having those consequences for people that are being measured through these tools.

**[Legal Framework]**

So what does the legal framework have to say about some of these? The good news is that the law does actually say quite a bit. I am pulling up here text of some key arguments in the Americans with Disabilities Act. I’m going to jump right to that, even though many analyses also look at Title 7 in the United States, which prohibits discrimination on the basis of race, gender, sexual orientation and ethnicity.

Within the Americans with Disabilities Act, there is actually some very clear and specific language that raises very real concerns about the legality of some of these tests and the potential discriminatory effects. The ADA requires that testing formats be accessible. We need to make sure that candidates can request reasonable accommodations. So that is one clear angle here — the test that I began showing which had the discerning four-colour dots, for example. An employer would need to be able to make a reasonable accommodation offer to someone that is unable to interface with the testing format.

But then the ADA says more. It has very useful language about prohibiting discrimination, including the prohibition of employment tests that screen out or tend to screen out a person with a disability. One of the obligations is on employers — is that they may be liable if they fail to select and administer employment tests in the most effective manner to ensure that tests accurately reflect the skills that they purport to measure, rather than reflecting an applicant’s or employee’s impairment. So there, what the law’s getting at is saying, is this test actually measuring what it is trying to measure? The skills that are required for this job, as opposed to evaluating someone’s ability or disability.

Another legal argument under the Americans with Disabilities Act is that there is a prohibition on pre-employment medical examinations. This comes up in settings, not exactly the test that I showed you in my opening remarks, but in other settings where there is an increasing use of personality quizzes and personality tests. There has actually been some good case law established that if the test draws on, for example, psychiatric evaluation measures, if it is administered by a psychologist, it crosses the line from being a hiring tool to being a medical examination that’s prohibited under the ADA.

So the good news is that there is some pretty strong language. The bad news is, it’s very hard to bring these claims. They rely on individuals knowing that they were the victims of discrimination, knowing how they were screened out, and then being able to effectively prosecute their rights. As we know, that can be enormously challenging. One of the things that we are very focused on is not only educating people about their rights, but then also going directly to employers so that they are thinking about the risks of exclusion pre-emptively and taking conscious affirmative steps to avoid them.

**[AI in Hiring]**

We have recently partnered with a group of over twenty civil rights organizations in the United States to publish the “Civil Rights Principles for Hiring Assessment Technologies.” This is intended as a tool for both policymakers, advocates and employers to know about the potential vectors of discrimination and take conscious steps to be able to address them. Some of the things that we emphasize in that is first of all a principle of non-discrimination, which should run through all of these tools.

The second one was a really important point for me and for my colleagues as disability advocates — this focus on job relatedness. Here, what we are worried about is the overwhelming tendency, a very common habit amongst employers, to think of tests and not actually go through that rigorous exercise of figuring out whether what they are assessing is essential for the job in question. That’s actually a requirement under the American Disabilities act, as I mentioned, but it is a very easy step to overlook, particularly when being sold a product off the shelf that purports to help you screen through candidates. But what we are saying is that essential principle needs to be the check of saying, what are the attributes that are actually required to do the job in question? How do I measure somebody’s ability to perform those particular skills required and nothing else? So that is an essential piece that we are pushing for very strongly.

The third is the goal of notice and explanation. So if somebody doesn’t know what they are being evaluated for or how they are being evaluated, it may actually be very hard for them to know whether their disability may adversely affect how they do in that test. And then it may be hard for the employer to know. For example, if you think about an employment test that is purporting to analyze how someone presents on a video, lack of eye contact may well have a significant impact on somebody’s success or failure under that metric. But an autistic person may not know that they need to ask for an accommodation or to flag that disability as a reason why they should have an alternative mode of testing. The notice piece here is essential. Again, I should pause to say I have fundamental concerns with the testing technology to begin with, so this is not to imply that I support that as a valid method of measuring someone’s effectiveness. But if we are going down the path of an employer knowing this tool, they need to very accurately and fully and fairly describe the process that they are using so that employees know how they may be affected in the course of that engagement.

The fourth principle that we are pushing for is rigorous auditing. It is very hard for people to know whether or not their platforms are discriminating against people, if they’re having adverse impact on particular groups, without rigorous and frequent checks of how people are doing and how they are coming through. For reasons I’ll talk about in a minute, I actually think that auditing is very hard in the disability context. Nevertheless, it is an essential principle that we need employers to be frequently checking and looking at the impact of the tools they are trying to use.

Then, the last point, number five, is about oversight and accountability. Here is the notion that we have to have employers understand their legal obligations, their ethical and moral obligations, if they are considering using these types of tools. And we need real oversight and accountability from lawmakers and from regulatory websites as well. In the United States, one area of focus is the Equal Employment Opportunity Commission. Another is looking at how federal contractors, which are a large percentage of the U.S. employer base, are using these tools and what accountability we can insist for there. And then the third is thinking about how legislators on Capitol Hill are engaging on these topics as well.

There was recently, in the United States, a bill introduced called the Algorithmic Accountability Act. What it mandates is that anybody using an algorithmic decision system has to audit for potential bias, and if they uncover it, have to take proactive steps to mitigate that bias. What’s interesting about that regulatory approach is that it doesn’t say exactly what the answer is. It doesn’t say, come down exactly on whether or not these tools should be used, but it says if you are making the choice, you need to have a certified impact assessment methodology that accompanies the use of that tool. That’s an important piece of the puzzle as well.

When I think about how to — where the points of intervention are, a lot of my work at the Center for Democracy and Technology is focusing, as I mentioned, on empowering advocates to know what potential areas of concern are, help raise general awareness. But there is also a lot of pressure to be put, as I said, on employers on their in-house counsel. So we are spending a lot of time talking to people at the American Bar Association and to other lawyers about what legal vulnerabilities look like in this space.

The third piece is thinking about how employers can be empowered to put pressure on vendors and to ask the right questions of vendors that are generating these tools. As I mentioned earlier in my remarks, it can sometimes be tempting to take a product off the shelf and to not think specifically about whether or not it’s right to use that product, how it should be tailored for your organization, whether it’s asking the right questions and actually screening for the things that matter for the role you are trying to fill. We want to empower employers and other decision-makers to be much smarter, more informed consumers of these products and make sure that they are engaging in that self-examination as to whether or not they are appropriate for use, and if they do use them, how to mitigate against potential concerns.

If you’ll indulge me, I’ll spend one more minute just talking about some of the challenges in the space and then I would love to pass it over to Julia and then to our broader discussion.

**[Challenges]**

On the challenges piece, there are a couple specific areas that come up in the disability space in particular when we’re thinking about the use of AI tools in hiring or in the workforce. One is that even if a company, a vendor of these types of products, is aware of the concerns, aware of the potential bias, fixing that at the design phase is actually quite challenging. Some of them are doing thoughtful outreach to groups to think about how we get more voices at the table to think about potential negative impacts of these tools.

But one of the challenges, and Julia may well talk about this, is that these tools train the way that AI learns is by taking an existing body of data, studying the patterns within it and from that, making inferences to informed decisions in the future. The sheer variety of ways in which disabilities may present, the sheer range of disabilities, coupled with the challenge of getting data about people’s disability, because many of us feel very concerned about sharing that publicly, let alone in a database that is used to train algorithms, the lack of training data makes it very hard for people that are trying to have their tools, understand the full range in which a person’s abilities may present during the course of an interview.

In a separate piece, which may well undermine the validity of these tools in general, is that when you think about basing decisions suitability for job based on trends you have seen in the existing workforce or in a selective training data set, by definition what you’re doing is making assumptions, drawing from a general crowd, as to what that means for a specific person. And that’s exactly what our civil rights laws and the disability rights movement has been fighting to stop for decades. A lot of those decisions and inferences are based on stereotypes and assumptions about people’s general abilities, as opposed to a decision that is based on somebody’s specific consideration of the individual candidate in front of you. So I think that’s a critical element that we need to think about when we are looking at the use of AI in hiring and thinking about the disability lens in particular.

My next point is a little technical, but I hope you’ll come on this journey with me, which are the limitations of auditing for bias on the basis of disability. Some people in the AI space, when we talk about the concerns of bias and hiring, say, well, one of the ways to mitigate this is to come in and do statistical auditing to see how people are doing in this test and see if we are inadvertently screening out women at a higher rate than men, for example. That’s not just a scientific approach, it actually is endorsed by the Equal Employment Opportunity Commission, which has on the books in its guidance on employee selection procedures which date from 1978, an idea that the best way to look for discrimination is to do a statistical analysis and follow what they call the four-fifths rule, which is that if a group is being screened out at four-fifths the rate of the dominant group. So if women are being screened out or making it through at a rate of only 80 percent how men are making it through, that is a good indication that there is something discriminatory happening in your tool and need to go back and re-evaluate it.

That model, that idea that you can screen for bias in a statistical way and then go in and use that to fix a tool is very dominant amongst the vendor of AI hiring tools right now. For many of them, they do follow that approach, and as a result, they then market their tools as being audited for bias and something that companies can comfortably use without concern. The problem is, that type of statistical auditing for bias is enormously hard when you think about disability. What data set would you be running through that in any statistical way would show how people are being screened out versus non-disabled people? When disabilities manifest in so many different ways and you often won’t have statistical significance in the number of candidates coming through. It’s much easier to do that type of statistical analysis for gender, for example, or for race. Now, it is equally problematic, I should say, for non-binary individuals or transgender individuals, for people of mixed race. But at least in the gender and racial diversity categories, we have more established categories and typically larger sample sizes that are easier to study.

So there is this real problem around auditing for bias on disability. The reason why that matters for all of us, as I said, a lot of vendors are out there at conferences, and I go there to these employment conferences, marketing their tools, saying that they have screened for bias. But they’re not thinking about disability when they make those pronouncements and that is hugely problematic from an inclusion perspective.

I’ll end on this last big point, which ties to that, which is just, as a general matter, I am so thrilled that this conversation is happening in this workshop because disability are too often excluded from discussions of algorithmic bias. We very frequently have seen this wonderful emergence of a conversation around algorithmic fairness, but it is still easier for people to jump to think about racial diversity and gender diversity, both very, very important, but not add in the additional ways in which people could be marginalized, and particularly when people over sect within those categories to be multiply marginalized. So it is an essential part of our program that we need to think about those crossover issues.

The final point is I think we need a lot more effort to think about how we move outside the bubble of policy advocates and academics that are talking about these issues to reach the decision-makers that are driving the market to create these tools and to deploy them. We need to be out talking to employers, talking to vendors, and trying to make smarter decision-makers throughout that process underscoring the ethical business and legal imperatives for change. Those are some of my opening thoughts. I’m thrilled to be here with you. Julia, I will pass it over to you.

**Julia Stoyanovich:** Wonderful. Thank you so much, Alexandra. This is a perfect segue. So, let’s get started. I am thrilled to be here. My name is Julia Stoyanovich. I’m an Assistant Professor of Computer Science and Engineering and of Data Science at New York University. I also co-direct the newly established Center for Responsible AI at NYU.

Today, it is my pleasure to speak with you about the responsible design development and use of algorithmic systems, particularly as they pertain to hiring, employment and the future of work, particularly, of course, as they pertain to the community of individuals with disabilities.

I have to say that I am humbled to be speaking in front of this audience, both because of the amazing strengths and creativity that members of this community have been exhibiting and resilience. And also because this is my first time speaking to a group of individuals that includes individuals with disabilities. I ask you in advance to please forgive me if I use incorrect terminology or if my presentation methodology or speed is inappropriate. I would love for you to correct me, to give me feedback so that I can become one of the advocates for this community in a thoughtful way.

**[Automated Hiring Systems]**

So our topic today is, of course, the future of work. Automated hiring systems do make a prominent portion of that topic, as Alexandra already discussed with all of us. Part of the discussion focuses on the use of technology for hiring, but I want to be sure that towards the end of my presentation, and also during questions and answers, we get to other topics that involve technology, and this community, and employment and the future of work. Things don’t stop at hiring.

To recap some of the things that Alexandra already said, in a recent report from Upturn, the hiring process is described as a funnel. It’s a sequence of steps in which a series of decisions lead to job offers to some individuals and rejections to others. This process starts when employers source candidates by ads or job postings. What I am showing on this slide is the depiction of this funnel, with the different stages shown pictorially.

Before I dive in, I guess I should say that many of the images here, the majority of the images, come from a scientific comic that I created together with the amazing Falaah Arif Khan. This comic is available online and it’s accessible through a screen reader. It’s currently released in English, and it will also be released in Spanish and also accessible in Spanish this week. So if you cannot see some of these images today, you can look at them later with annotations in the comic book.

So, this funnel that I am depicting is made up of a sequence of steps. These are sourcing of candidates by ads or job postings. The next stage is typically called screening, where employers assess candidates by analyzing their experience, skills and characteristics. Next, through interviewing, employers continue their assessment more directly. After that, background checks may follow. And then, during the selection step, employers make final hiring and compensation determinations. Importantly, data and predictive analytics are used during all of these stages. As stated by Jenny Yang, former commissioner of the U.S. Equal Opportunity Commission, or EOC, automated hiring systems act as modern gatekeepers to economic opportunity.

**[And Now … Some Bad News]**

And now, of course, we have some bad news, as Alexandra already spoke about things that can go wrong in quite some detail. So I will just recap very quickly here that we have been seeing cases of discrimination based on gender, on race and on disability status at all stages of this pipeline. The kind of concerns that we are faced with the use of these automated hiring tools pertain to unfairness in the decisions that are made by these systems, which we would denote as discrimination, or by the unfairness in the process by which these decisions are made. And these I would call due process violations or due process concerns.

Very often, discrimination and due process violations are linked to the term “bias.” This is a term that some of you already started discussing in chat that we will revisit momentarily. There is also concern, before we dive, into bias about whether these tools actually work. So what I’m showing here on the slide is a depiction of Arvind Narayanan. He is a computer science professor at Princeton who has a wonderful talk about spotting AI snake oil. So, are these tools actually working? Are they picking up useful signal from the data or are they an elaborate coin flip at best? As Arvind Narayanan puts it, “Are these tools AI snake oil?”

In the complex ecosystem in which automated hiring tools are commissioned, developed and used, we must ask ourselves who is responsible for ensuring that these tools are built and used appropriately. Who is responsible for catching and mitigating discrimination and due process violations, and for controlling the proliferation of AI snake oil under the fancy label of data science and AI? What I am showing here, in support of this narrative, is a culpability lineup, in which the point that I am making is that everybody is responsible — scientists are responsible, members of the public, platforms, software developers, all of us are responsible for making sure that these tools are used to benefit society and not harm individuals or particular groups, such as individuals with disabilities.

**[Automated Decision Systems (ADS)]**

The hiring funnel, as well as each component of the funnel, are examples of automated decision systems or ADS. These systems process data about people, some of which may be sensitive or proprietary. They help make decisions that are consequential to people’s lives and livelihoods. They involve a combination of human and automated decision-making. They are designed with the stated goals of improving efficiency and promoting, or at least not hindering, equitable access to opportunity. And finally, they are subject to auditing for legal compliance and, at least potentially, to public disclosure. ADS, automated decision systems, may or may not use AI; although, most of them are billed as AI because AI sells. And they may or may not have autonomy, meaning they may not be making these decisions entirely on their own. Usually, there is also a human decision-maker in the mix, but they all rely heavily on data. So what I would like for us to focus on is the role of data in this environment.

**[Regulating ADS?]**

In response to the question about responsibility, we’ve been seeing attempts to regulate the use of data-driven algorithmic tools, such as those that make part of the hiring funnel. This activity is broader in scope than algorithmic hiring, of course, and it pertains to automated decision systems, more generally, the systems that I described previously. A big question here is: How might we go about regulating these systems? How might we be going about regulating ADS? And should we even attempt to do this?

And while the predominant sentiment in the industry is still that regulation will stifle innovation — I am showing here a reckless child on a bicycle not hitting the brakes — industry alone doesn’t get to decide. Even in the Silicon Valley, the need for meaningful regulation to ease compliance and to limit liability is starting to be more and more broadly recognized. There is much debate on a specific regulatory framework that we should adopt. Should we use precautionary principles that can be summarized as better safe than sorry — here I am showing an image of a child on a bicycle in protective gear — or, more likely, attempt a more agile, risk-based method, such as algorithmic impact assessment — here I am showing a picture of a child who is riding a bicycle carefully with a helmet. All this, and more, is the subject of intense debate. Some of which, I had a chance to witness firsthand and in which I am still actively participating.

**[ADS Regulation in NYC: Take 1]**

New York City, where I live, recently made a very public commitment to opening the black box of New York city government’s use of technology. In May 2018, an automated decision systems task force was convened, the first such in the United States, and charged with providing recommendations to New York City’s agencies about becoming transparent and accountable in their use of the use of ADS. I was a member of this task force by appointment from the New York City Mayor.

The task force issued a report that included a set of principles. These principles say that we should be using ADS only if they promote innovation and efficiency in service delivery, not simply because they are available, not simply because they are being sold to us, as this AI snake oil. We should also be promoting fairness, equity, accountability and transparency in the use of ADS. Finally, we need to be thinking about how to reduce potential harm across the entire lifespan of ADS, starting from its design and all the way through deployment.

While making important points in this report, the ADS task force unfortunately didn’t go very far in terms of concrete recommendations. I am happy to discuss the reasons for this offline, but I can spend hours talking about what went right and what went wrong during our deliberation. But we also have an immediate opportunity to make things a lot more concrete in the context specifically of regulating hiring systems in New York.

**[Regulating Hiring ADS: Int. 1894–2020]**

There is currently a proposed law, a bill, that is being considered by the New York City Council Committee on Technology. There was a hearing on this bill, just this past Friday. This bill would regulate the use of automated employment decision tools with the help of a biased audit, as well as public disclosure. Individuals being evaluated with the help of algorithmic tools would have the right to know that an algorithm, rather than a person, was used to screen them. Most importantly, in my mind, individuals would also be told what job qualifications or characteristics were used by the tool. Ideally, we would also want to make sure that the job relevance of these qualifications is substantiated.

**[Framing Technical Solutions]**

So let me now switch gears and talk about the things I was asked to discuss, and that is, what is this bias in the data? What are some of the technical interventions, some of the technical solutions that we may want to use here?

To start, I want us to step back and think carefully about the role of technological interventions, such as data and model bias, that you may have heard about. This discussion is necessary to help us find a pragmatic middle ground between two harmful extremes. One of these is techno-optimism. What I’m depicting here is an image of a woman with dark sunglasses. In the sunglasses, there is a reflection. On the left, I am reflecting one harmful extreme and that is techno-optimism, a belief that technology can single-handedly fix deep-seated societal problems, like structural discrimination in hiring. On the right, I am showing an image that illustrates techno-bashing. That is a belief that any attempt to operationalize legal compliance and ethics in technology will amount to fearwashing, and so should be dismissed outright. Our job, I think, is to really find a way to navigate between these extremes to create an nuanced understanding of the role of technology in society.

**[“Bias” in Predictive Analytics]**

So let’s get back to bias, a term that is used very often these days to explain what’s wrong with automated decision systems but that remains poorly understood. What do we mean by bias? Our meaning of this term is not in the traditional sense that is used by statisticians, who say that the model may be biased if it does not summarize the data correctly. Instead, what we are seeing here are examples of societal bias exhibiting itself in the data. Let’s unpack that further.

**[Data, A Reflection of the World]**

Data is an image of the world, it’s mirror reflection. When we think about societal bias in the data, we interrogate this reflection. This is what I’m depicting here, a world on the left that we don’t quite know, and then there is a person that is looking at this world through a mirror, through a lens.

One interpretation of bias in the data is that this reflection is distorted. We may systematically oversample or undersample, overrepresent or underrepresent particular parts of the world, or we may otherwise distort the readings. It is important to keep in mind that the reflection cannot know whether it is distorted. In other words, data alone cannot tell us whether there is a distorted reflection of a perfect world, a perfect reflection of a distorted world, or if these distortions compound. The assumed or externally verified nature of the distortion has to be explicitly stated.

Another interpretation of bias in the data that I’m showing on the right of the slide is that, even if we were able to reflect the world perfectly in the data, even if we were able to take a perfect measurement of the world such as it is, it would still be a measurement of the world such as it is or such it has been historically and not as it could or should be. Once again, importantly, it’s not up to data or algorithms, but rather up to people — individuals, groups, and society at large to come to consensus about whether the world is how it should be or if it needs to be improved. If so, how we should go about improving it.

**[Changing the Reflection Won’t Change the World]**

My final point on this metaphor is that changing the reflection does not necessarily change the world. If the reflection itself is used to make important decisions, for example whom to hire or what salary to offer to an individual being hired, then compensating for the distortions is worthwhile. But the mirror metaphor only takes us so far. We have to work much harder, usually going far beyond purely technological solutions to propagate the changes back into the world, not merely brush up the reflection.

**[“Bias” in Predictive Analytics, Revisited]**

One way to conceptualize automated decision systems is simply as a so-called predictive analytic that takes as input a nice, clean rectangular dataset that I’m depicting here, it crunches it, then it produces a result. This result could, for example, be a prediction of how likely somebody is to do well on a job, and so therefore, whether we should be hiring them. If we don’t notice that the result is such that no women are shown ads for high-paying jobs, for example, and no individuals with disabilities pass an online job interview, then we have three choices, if this is our worldview. And these are that we could tweak the input data; for example, upsample or downsample some groups. We could tweak the algorithmic box that crunches this data or we could change the result; for example, we could reassign outcomes.

However, an issue here is that this particular view is a frog’s-eye view. I argue that we need to expand this quote, and also, think, what specifically happens inside this box that crunches the data, think how the results are being used — Whom are they impacting? Who benefits and who is harmed? And also, we need to ask ourselves, where did the data come from? In other words, we will gain much more power to incorporate responsibility into automated decision systems development and use if we see these systems through the lens of their development, design and deployment lifecycle, and their data lifecycle.

**[Bias in ADS]**

In their seminal 1996 paper, Friedman and Nissenbaum identified three types of bias that can arise in computer systems with the broader view of the systems. This bias is represented here as a three-headed dragon: pre-existing, technical and emergent. Pre-existing bias exists independently of an algorithm itself; it has its origins in society. This is the societal bias that we have been discussing with the mirror metaphor. Technical bias can be introduced at any stage of the system’s lifecycle and it may exacerbate pre-existing bias. I will give a couple of examples next.

Finally, emergent bias arises in the context of use, and it may be present if a system was designed with different users in mind or when societal concepts shift over time. In hiring, a prominent example of this is “the rich get richer.” Emergent bias arises because decision-makers, hiring managers in this case, tend to trust the algorithmic system to, indeed, select the most suitable candidates, for example, by placing them at top positions of a ranked list. This, in turn, going to shape a hiring manager’s idea of what the suitable candidate looks like.

**[Models and Assumptions]**

To discuss technical bias a bit more, let us look for about a minute at models and assumptions. Technical bias often has its origin in incorrect modelling ad assumptions or in technical choices that follow from these assumptions.

Here are some concrete examples. What I’m depicting here is an art gallery in which there are four paintings of apples: one is the realistic one and three are abstract to different levels of obstruction. Suppose that the job applicant applies through an online form and that this form allows applicants to leave their age unspecified. This data will travel through multiple data processing steps before it’s given as input to a predictive analytic, a classifier that will decide whom to invite for a job interview. Further, to make a prediction, the classifier will need to know what the value of age is. So if that value was unspecified. it will try to guess it, to interpolate it, to fill in the blank. Now the question is how should the system guess someone’s age. The most common method for this is the simplest one; it is called mode imputation, replacing a missing value with the most frequent value for the future. If age is missing at random, meaning that everybody is equally as likely to not specify their age, then this method is appropriate. But if age is missing more frequently for older individuals, for example, because they may fear age discrimination in employment, then mode imputation will impute age for them incorrectly, and systematically so.

Next, consider an online form that gives job applicants —

[Brief pause to address technical issues.]

Restarting with the Apple gallery, we talked about modeling assumptions actually being extremely important, and we said that if you guess a person’s age or an age of a demographic group systematically incorrectly, then this is problematic and it is based on some assumptions you are making about your data.

The next example I was going to give is a form that gives job applicants a binary choice of gender but also allows them to leave gender unspecified. Suppose that about half of the users of the attribute identify as men and half as women, but that women are more likely to omit gender because, once again, they fear discrimination, perhaps. Then, if replacing a missing value with the most common value for that attribute for gender is used, then all predominantly female unspecified gender values will be set to male. More generally, multiclass classification for guessing values, missing values for missing value imputation, typically only uses the most frequent classes as target variables, leading to a distortion for small population groups because membership in these groups will never be imputed.

Next, suppose that some individuals identify as non-binary. Because the system only supports male, female and unspecified as options, these individuals will likely leave their gender unspecified. Then, once again, the system will use mode imputation and it will set their gender to some value, probably male, because this is a predominately male dataset. A more sophisticated imputation method exists, right, and it could be used and it could do better, but it will still use values from the active domain of the feature as we see, meaning that it will still set the value to either male or female. This example illustrates that technical bias can arise from an incomplete or incorrect choice of data representation.

Finally, consider a form that has home address as a field, and a homeless person may leave this value unspecified, and it would be incorrect to impute it. We would lose information. While dealing with blank null values is known to be difficult, and these already considered among the issues of data cleaning, that this is a technical discipline that thinks about these questions, the needs of responsible data science introduce new problems here, with much higher stakes. Importantly, it has been documented that data quality issues, including missing values, often disproportionately affect members of historically disadvantaged groups, such as individuals with disabilities. So this is a risk that we see here of technical bias exacerbating pre-existing bias for such groups. There are, of course, immediate parallels here for individuals with disabilities of their data being missing, missing in a way that is non-random, and imputation potentially being harmful if it’s not done thoughtfully.

**[Dimensions of Data Inequality]**

I’m going to skip this. This is the framing that I like that talks about all of these problems under the lens of data equity representation, feature, outcome and access.

**[Making ADS Work for *All* of Us]**

But what I want to stress is that, rather than focusing on the use of AI (artificial intelligence) and automated decision systems for hiring and screening, we should instead shift our focus to how we can use AI to solve problems that create opportunities for all of us. Rather than developing platforms that “encode inhospitality,” a term due to Chancey Fleet, we should be thinking about how to use AI to improve accessibility. Rather than ghostwriting code, again a term due to Chancey Fleet that enables accessibility, what about making writing that code part of a tooling of a toolkit that has our priority, to which we pay particular attention. What about developing AI methods that are specifically working for individuals with disabilities? I am showing here a picture of Chancey Fleet and a quote from one of her talks.

I am going to skip this; this is a picture that illustrates how individuals with disabilities can be seen with even more errors by facial recognition software which, frankly, we are in more trouble if that software works than if it doesn’t.

**[Take-Aways]**

My takeaways are that we should be thinking about how to build technological systems that are rooted in the needs of people. This means that we need to expose the knobs of responsibility to people. We must work together to create meaningful regulatory mechanisms and also engage in educational activities. These are the goals of the Center for Responsible AI at NYU that I am starting, together with colleagues and friends.

An example of an ongoing educational activity is a series of workshops that will culminate in a course for members of the public on the use of algorithms in general and, specifically, their use in hiring and employment. We are doing this together with the Queens Public Library in New York City as a pilot, and will then scale up through libraries in the U.S. Another example is a comic series of which we have released the first volume, called “Mirror, Mirror,” together with the very talented Falaah Arif Khan, who I think participated in an earlier workshop that is part of this series. So please take a look at the comic; let us know what you think.

The final anecdote which I will leave you is, I already mentioned, that the comic is accessible, you can read it with a screen reader. The only way that we were able to make this work is by getting feedback from Chancey Fleet. She helped us debug it. And, of course, that method doesn’t scale, so another opportunity for using AI is to help us make educational materials accessible, including also on AI itself. I will stop here. I am happy to continue the discussion. Thank you very much.

**Vera:** Thank you very much. Thank you, Julia. Thank you, Alexandra. That was two really, really terrific presentations. I really felt that I was learning so much and I could tell that — I was keeping an eye on the chat throughout both of your talks — and I could see that people were quite engaged and had some really important questions to ask.

I think that one of the difficult challenges with AI is that notion of that black box and understanding how we can get in it and get at it and write policies related to it. I thought that you both really helped to open that up for me a little bit today.

Alexandra, when you were talking about policy, particularly with regards to the ADA, I was trying to think about how that’s playing out in Canada and other jurisdictions. I wondered, have you had opportunity yet to focus on any other jurisdictions in terms of regulatory structures around or protections around the use of automated data systems, decision-making systems, and hiring processes or other areas? I thought it was well laid out in Title I in the ADA. We have similar human rights codes here in Canada that protect against discrimination, but I wondered, is that something you’ve seen any other areas where they have got some novel approaches to policy?

**Alexandra:** Sure, I felt self-conscious joining this workshop when I don’t have familiarity with Canadian law, so you have to excuse me with my U.S. focus. We are spending a lot of time in Europe, as well. So there, the European Commission is focused on questions of AI and equity and bias. The European Union moves slowly but steadily when they decide to regulate a space, and we saw that with their general data protection regulation, which was like a slow-moving glacier, but did end up having a significant impact on privacy policies around the world. They have some provisions on AI in that bill and they’re now considering a new initiative around AI, as well.

Part of the problem though is that a lot of these regulatory regimes are still just focused only on transparency and putting the burden on the user. There were some comments about this in the chat. In my mind that is asking way too much of all of us as individuals, to know how we may be discriminated against and be able to take individual actions. So that’s why we’ve been focused so much on trying to get directly to employers to know here’s the legal framework, here’s how you may be violating the law, but also just here from a moral and ethical perspective, is why you need to focus on these things.

The last piece I’ll say is Christopher had excellent comments in the chat about how, even in the U.S., that language in the American Disabilities Act is really, really good. I would argue it is basically undiscovered, so far as it comes to AI, because nobody is actually litigating these cases. The Bazelon Center for Mental Health Law did bring a series of challenges in the early 2000s around the use of personality tests. They filed complaints with the EEOC and fought for years. And even there, they had a hard time moving the needle, despite the new language on the books. So that tells me a couple of things. One, we need to keep stepping up our advocacy and our pressure. Two, we do need to look at this legal framework and see what else we could be doing. So there is a new administration coming in the United States, could that new administration issue new guidance from our Equal Employment Opportunity Commission? To say, actually, this is, indeed, a violation of federal law and lean in far more aggressively to help set expectations on what is acceptable testing and what is not.

**Vera:** Thank you. It’s so interesting for us to hear about what is happening elsewhere, including, and really, don’t feel self-conscious about having an American focus, we’re well aware of where you were coming from when we invited you to be here. We’re interested in what is happening and in other approaches from other jurisdictions, in part because we are wanting to consider what might be some possible approaches within our study group here today. So getting a variety of information is really, really quite helpful to us in this.

**Julia:** If I may add to this, we actually are looking to Canada also for some help and some guidance on this. I pasted in the chat the link to the Canadian Federal Directive on Automated Decision-Making. We don’t have such a directive in the United States. It takes an approach to regulation, this is the use of the tools within the Canadian government, not more broadly, so regulating the hiring ecosystem is out of scope, but the approach that they take is based on impact assessment, which in my mind is the right way to go, so understanding what are the potential harms, what are the benefits, and to whom are these harms and benefits. One thing that is still lacking, even if you pass regulation of the sort, you must have an informed public. And I think this is really where the rubber hits the road. We all need to think very carefully, and none of us know how to do this, not in Europe, not in New Zealand, not in Canada, not in the U.S., how to actually create even a basic level of understanding of data, algorithms, and what we should be asking of algorithms and of the data. So this is my two cents as a technologist. I think that technology is the easy part here, relatively speaking. It’s really educating.

**Vera:** It was fascinating when you’re talking about the notion of whether it’s a distortion of a perfect world, a reflection of a distorted world, and these different ways of understanding the data, and I think there was a lot that was very interesting. It gave me new ways to think about this whole process.

I will say, it struck me, I think, Alexandra, you were the one that was talking about the idea that, really, when the systems are being used, for example, in the hiring process that there should be disclosure, that we are being put to the test or put through certain systems. I know that even for myself, when I’ve looked at some processes, I’ve wondered, are there hidden buzzwords I should be hiding? White font in my resume or my CV, or if I’m helping someone, I’m wondering if that is something I should do. There are these hurdles that are there that many of us are not aware of, no matter who is looking for the job. I can certainly see how that idea that this should be disclosed is — I think might be novel to some employers who sort of like the shield of these systems between them and the hiring process, them and the candidate.

And then, of course, some people, maybe without intending to, think that because it is a data system that it will not have bias, that it is going to be unbiased because there is no human in it to bias it. But, of course, one of the things we are learning and are exploring is the fact that bias is inherent in the systems because it is built by biased people and by biased data. So addressing that is really — the solution you were suggesting, Julia, was fascinating to me.

I think I’m going to try to talk less. [Audience member], you had quite a number of suggestions. Why don’t you choose one of your favourites to start with, and we can come back to you again after a couple others, if we have the opportunity.

**Audience Member:** Great. I have one question and then a couple points I’ll leave, in case you happen to address them or not. The question is around the disclosure piece. How do we monitor for discrimination based on disability when disclosure should be optional and discrimination based on disability should not be predicated on a required disclosure? So that’s the primary question I wanted to ask. I’m also curious to know if there are easy flags to watch for, for discriminatory algorithms; for example, if gaps in job history could be a bad thing that screeners are used to look for, or if there are really good flags, like are they looking for the benefits of lived experience? Easy ways to be able to tell who is likely to be a real problem versus somebody who is going to get — so I guess that’s two questions.

**Alexandra:** They are really good ones, and ones that we have been spending a lot of time thinking about, because again, in theory, the Americans with Disabilities Act helps on this. Your request for accommodation cannot be used against you. But then we have to dovetail that reality of the legal framework and people’s actual experience and their level of comfort asking for an accommodation, and realize just how challenging that is.

A system that requires disclosure in order to be fair is really an unfortunate one, and one that I think it quite challenging. For the reasons I mentioned in my talk, it also makes the data collection piece really hard. Even if you wanted to come in as a good-faith researcher and audit from behind how a sample of an employee’s did, you could maybe make inferences about gender or race based on what you know about a person and do a loose assessment, at least, of how people tend to be trending from a risk of discrimination on those vectors. Disability, much harder, there really does need self-disclosure for people to know for auditing purposes how you were treated.

We’ve been having some conversations on whether there are creative solutions here. Could we have a co-op of disabled people who voluntarily agree that they are going to share their data with a trusted third-party non-profit and then go out and apply for jobs and see how they’re doing? But the mechanism of that would be a lot of working out. There are concerns about that. I’m a privacy lawyer in addition to thinking about fairness. But I do wonder some of those things. I almost wish that LinkedIn had a way of you opting into a voluntary program where you could reveal your protected characteristics, just for purposes of trusted auditing — that the employer wouldn’t see them but a trusted auditor would. That’s one of the thoughts I have at night thinking about that question, but I think we need smart minds really coming together on that point.

As for the flags, I think there, this is why we need to know better what people are being evaluated on so that you can look to think through what the flags are going to be. Gaps in a resumé might be one obvious one, but there may be many other ways that a system may discriminate against a particular person with a disability based on their individual circumstances. Data scientists can’t even think about or contemplate to know about a potential risk. So I flagged in my talk some of the most obvious ones that jump to me. I tried to spend a lot of time cataloguing these from talking to folks who are experiencing some of these tools, but t needs to be an ongoing project, again because of the sheer diversity of people’s experiences that people will have when interacting with these tools.

**Julia:** If I may add to this from a technological point of view, I absolutely agree with everything you said, Alexandra. There are actually technical methods that can help us support these types of auditing as well as these types of public disclosure. When we talk about auditing, it doesn’t have to be us measuring performance of a system on the data, the life data on which it is deployed. We have methods to generate data synthetically, or semi-synthetically, that resembles the kind of population on which we want to interrogate the behaviour of this system. So that technology is there. It can give us privacy protection and it can allow us to work with these machines to try and test them until we are satisfied with their performance along all of these dimensions.

I also want to add that, in addition to biased data and biased people, an important dimension here is that the criteria of goodness against which we evaluate these systems can be biased as well, because there is no such thing as absolute perfect accuracy. If you are accurate for one group, you may be less accurate for another, so who decides on that trade-off? Who decides what accuracy means, what fairness means, technically? So that’s kind of another dimension there. In terms of job relevance, that’s crucial. It’s not enough for us to just say, the reason you were selected for this job is because your name is Jared and you played lacrosse. This may, in fact, be, and this is an anecdote that’s well known, the signal that the system picked up. But if we know that this is the signal that the system picked up, we should immediately challenge it, right?

So what I think is a promising metaphor here, technically. So for auditing we had synthetic data sets, but for understanding which features are job relevant and which features we think are okay to use is this metaphor of a nutritional label that I’ve been running with. When you explain to a particular job applicant how a decision was made that affects them, you can tell them it’s because your speed of discriminating between the red and the green circles was lower than what I expected. Then the individual would know that this actually is because of a particular disability that they are leaving with, and then they can challenge that decision. So I think we need, in addition to this auditing that talks about the behaviour of an algorithm in bulk over an entire population, very powerful mechanisms to explain to individuals what happened, who can then challenge these decisions.

**Alexandra:** Julia, I love that. One point I would add that complicates these is, for many of the tools, the value they’re selling is that they may not actually know what characteristics they’re evaluating for. They are studying trends in an existing pool of employees who are doing well, and then saying, they play this game this way, and so we’re going to find you people who play the game the same way. And they may well not be able to articulate the specific job characteristics that are actually making these people look like winners.

So our methods, what we’re saying is, let’s challenge that and say, no, you have to be able to articulate what attributes you are evaluating and what rate you are evaluating them so we can see whether those are tied to the essential function of the job. The problem is that actually cuts right to the business proposition of why we need this magic AI, the snake oil to be doing that inference for us, right? That’s going to be a struggle with the industry that we need to keep pushing for.

**Julia:** But we must. We absolutely must. If we don’t do this now — it may already be too late, in fact — but the longer we wait, the closer we will be to the state that we are in with regulating our inability to regulate social networks, for example, and the kinds of effects that we’re seeing there, that the only way to cancel these effects, really, is to dismantle the business models, of advertising targeting, in that case, that also hits us here as well because as targeting for jobs, as part of this funnel for hiring.

**Vera:** Right. Any other questions? I am just looking for any raised hands. I want to give people a moment to find that. This is the time to disrupt the system, absolutely. I couldn’t agree more. [Audience member] has put a question up.

**Audience Member:** Yes. I think others have heard me speak about this before. In one way or another, we’ve had education at some level. Can you here me?

**Vera:** There’s a little bit of distortion, but we are hearing you.

**Audience Member:** Okay, so I will speak slowly. How about people with disabilities, the nature of the disability or their social-economical identities, the location, their poverty, even, may not necessarily qualify to apply for the jobs that show up in highly technical apps or systems or platforms, right? I’m thinking, I know folks with disabilities, with Down syndrome. [Inaudible.] They work on the folks who are not as educated, who are a lot less disadvantaged than us here in this very privileged circle.

**Alexandra:** Thank you so much for that point. I have two thoughts to add. Other workshop members, I’m curious if they have thoughts. One is, just as a matter of course in our advocacy materials, we are now creating plain language versions of everything we write. And Know Your Rights versions of documents to help empower people to better understand these issues. Julia, I love that you have a comic book that’s another way of trying to make these issues accessible. Ours are literally a single piece of paper that just try to spell out very simply how these systems might affect you and what you can do about it.

The other point I heard in your question is realizing that a lot of these hiring tools right now are being used more for white collar jobs, we would say. So what do we think about other aspects of the workforce? One is that the personality tests that I mentioned are being used very widely. So they are being used in many retail jobs, in particular. Your Walmarts, CVS, Best Buy, Target are using those types of tests and have actually taken some steps in response to accusations of bias that come from those personality tests. So that’s one area where we do need to be vigilant.

The other is that we need to think about the use of AI, not just in hiring but what it means to be a worker that is subject to surveillance. Performance is being measured by tools. There is being very good research done about what this means for truckers, for example. Long-distance truckers who now are some of the most surveilled workers in the country, with machines looking at their eye contact and where their eyes are looking. For people that work in the Amazon warehouses, for example, Uber drivers. There are many examples of the quantified worker and what it means to have your movements tracked. And not only tracked, but evaluated against peers, and you rise or fall based on how you’re doing compared to that large dataset of others. What that means for human dignity, what it means for fairness, and how you’re evaluated and judged. So that’s another big piece of the puzzle that I think the conversation needs to go to next.

**Vera:** Yes, in fact, I was thinking we tend to look at the hiring process often, but the whole employment cycle, which both of you spoke about, is also that area that we need to consider and focus — the surveillance aspect that you’ve talked about, which is creating other datasets. It’s also information that’s being collected about one aspect of your performance or what you’re doing, but then maybe used in other ways, which is usually an ethical no-no, right, to collect data for one purpose and then use it for other purposes is not usually okay. Oh, EEOC, that’s your Employment Equity, someone wants to know what that means.

**Julia:** It’s the Equal Employment Opportunity Commission where Jenny Yang was commissioner under Obama, and we hope that she will now become active again in some capacity.

**Vera:** We are waiting with baited breath to see how things will resolve in your country for your next administration.

**Julia:** If I may just add one more very quick point on the discussion before, and that is the final point with my mirror metaphor is that you can’t just touch up the reflection, right? It does help to hire people who are qualified for the job, despite them maybe looking not as good on paper. But it’s also not enough for us to actually be mitigating at various steps of the hiring funnel where, generally, at the level of this hiring funnel is not enough.

What we need to do is to use AI and our human, personal thinking in figuring out how to make it so that people from the disability community, those who are severely disabled, learn the skills that are necessary to enter the job market. For this AI can and should help. I think that the future of work in this AI context is not, how do we make it so that we can use AI to hire fairly? Rather, it’s how do we make AI serve our purposes? Create new jobs for us. What are these jobs that AI can enable? How can we make AI make education more equitable?

**Vera:** I love the fact that you’re looking at the opportunity aspect. I think that we try to catch ourselves, as well, in not assuming that all AI is somehow problematic. For sure there’s lots of challenges, but if we can harness some of the benefits, that would be really terrific.

We’ve come to the end of our slotted time and I know there’s always still some questions. Really, I think I speak for everybody here when I say we really, really enjoyed hearing your presentations and have been the opportunity to have some questions with you. Thank you so much for joining us today. Thank you for all of the study group and our other participants who joined us today. We will continue our discussions in Canvas and elsewhere. We always hope that we’ll have an opportunity to engage with our panelists again, so Julia, Alexandra, hopefully, we’ll have more opportunities to share the idea around data trust, for example, something we’re exploring really deeply here at the IDRC. I see so much overlap in both of your work with interest here. We hope to see you again and collaborate further. Thank you very much for joining us today.

**Julia:** Thank you we’d be delighted to work together.

**Vera:** Everyone, if you have a reaction, you can put your applause in the windows. That concludes our session, so thank you very, very much. Much appreciated, everybody’s participation today.

**Alexandra:** Thank you very much.