Episode 6: Responsible AI and Machine Learning

One of the hottest topics out there is the rise of artificial intelligence and in particular of generative AI tools such as ChatGPT. Suddenly machines can pass MBA exams, write memos for you, create computer codes, and much much more. At the same time, the conversation around the potential harms, and in particular the inequities, that can be created by such technologies is also accelerating. How can we take advantage of all of the incredible things it can do without suffering from its potential harms? And, is that even possible? In this episode, we talk to Annie Veillet and Allison Cohen to answer these questions.

Featured guests:

Allison Cohen is Senior Applied AI Projects Manager at Mila. In this role, Allison works closely with AI researchers, social science experts and external partners to professionalize and deploy socially beneficial AI projects. Her portfolio of work includes: a misogyny detection and correction tool; an application that can identify online activity that is suspected of containing human trafficking victims; and an agricultural analytics tool to support sustainable practices among smallholder farmers in Rwanda. She was on InspiredMinds! Top 50 Influential Women in AI list and was the Runner Up for the 2022 Women in AI "Leader of the Year" Award in the category of Equity, Diversity and Inclusion.

Annie Veillet is a partner at PwC and leads their AI and Intelligent Automation offerings. She brings extensive experience in the planning and execution of complex Advanced Analytics and Automation solutions. Her experience includes Analytics transformations (from the identification of business drivers to implementing and scaling capabilities), process automation (with embedded AI components), customer segmentation using behavioral traits, and data management. Annie co-led the development of PwC's *Global Responsible AI Toolkit* and was recognized as one of the Top 30 Influential Women Advancing AI in Montreal by RE.WORK.

Moderator: **Brian Keng** is a Research Director at Borealis AI, a research centre created by Royal Bank of Canada, and an Adjunct Professor in Data Science at the Rotman School. At Borealis, he leads the incubator program building out innovative AI-enabled products and capabilities for the financial services industry. At Rotman, he plays a key role in shaping data science education and research through his work with the TD Management Data and Analytics Lab and the Master of Management Analytics program. Brian's primary professional focus revolves around building scalable AI and machine learning systems to provide robust solutions to core business problems.

Resources:

- Check out GATE's report on An Equity Lens on Artificial Intelligence
- Gender Analytics: Possibilities conference

Gender Analytics is a way to analyze your products, services, processes and policies with a gender lens to uncover hidden opportunities for innovation and improved effectiveness by considering gender, race, Indigeneity, disability, ethnicity, sexual orientation and other identities. Learn more here: https://www.gendereconomy.org/gender-analytics-online/

Want to hear more from the Institute for Gender and the Economy? Check out our signature podcast series, Busted, which busts prominent myths about gender and the economy!

Credits:

Produced by: Sarah Kaplan

Edited by: Ian Gormely

Transcript

Allison Cohen: You're asking a model to come up with some synthesized view of the world. And by definition, the data that you're training it with is noisy. It's on this side. It's on that side. It's up here, it's down there, and you have to come to some middle point and it ends up being a real abstraction of the reality. And so, the nuance that exists on the web is completely wiped away and any diversity that you were hoping for is gone. And where they converge is the area that's statistically significant in the data. So, those that are over represented already, those that are mainstream, are the voices that end up being reproduced.

Sarah Kaplan: One of the hottest topics out there is the rise of artificial intelligence and in particular of generative AI tools such as ChatGPT. Suddenly machines can pass MBA exams, write memos for you, create computer codes, and much much more. At the same time, the conversation around the potential harms, and in particular the inequities that can be created by such technologies is also accelerating. How can we take advantage of all of the incredible things it can do without suffering from its potential harms? And, is that even possible?

Welcome to episode 6 of Designing for Everyone, a podcast by the Institute for Gender and the Economy or GATE. I'm Sarah Kaplan (she/her pronouns), and a Professor of Strategic Management at the University of Toronto's Rotman School of Management, Founding Director of GATE, and your podcast host. In this 7-part limited series, we're featuring a high-impact set of conversations we had in April 2023 at our Gender Analytics: Possibilities conference.

To get into a discussion on AI and inequality, we invited 3 experts:

Annie Veillet is a partner at PwC and leads their AI and Intelligent Automation offerings. She brings extensive experience in the planning and execution of complex Advanced Analytics and Automation solutions. Her experience includes Analytics transformations, from the identification of business drivers to implementing and scaling capabilities, process automation with embedded AI components, customer segmentation using behavioral traits. Annie co-led the development of PwC's Global Responsible AI Toolkit and was recognized as one of the Top 30 Influential Women Advancing AI in Montreal.

Allison Cohen is Senior Applied AI Projects Manager at Mila, a community of scientists and interdisciplinary teams committed to advancing artificial intelligence for the benefit of all. In this role, Allison works closely with AI researchers, social science experts and external partners to professionalize and deploy socially beneficial AI projects, including a misogyny detection and correction tool; an application that can identify online activity related to human trafficking; and an agricultural analytics tool to support sustainable practices among smallholder farmers in Rwanda. She was on InspiredMinds! Top 50 Influential Women in AI list and was the Runner Up for the 2022 Women in AI "Leader of the Year" Award.

And, their conversation was moderated by Brian Keng who is a Research Director at Borealis AI, a research centre created by Royal Bank of Canada, and an Adjunct Professor in Data Science at the Rotman School. At Borealis, he leads the incubator program building out innovative AI-enabled products and capabilities for the financial services industry. At Rotman, he plays a key role in shaping data science education and research through his work with the TD Management Data and Analytics Lab and the Master of Management Analytics program. Brian's primary professional focus revolves around building scalable AI and machine learning systems to provide robust solutions to core business problems.

They discussed important ways that AI potential can be achieved without reinforcing inequalities.

Brian Keng: So I thought where we start with this is to just talk a little bit some of the projects you've been working on, which is very relevant to the conference. Maybe we'll start with you, Annie. I know you co-led the development of PWC's Global Responsible AI Toolkit. Can you maybe talk about how the impacts been to your customers and maybe some of the business that has led to?

Annie Veillet: Sure, absolutely. So, first, before I get into the impact, just give you guys a brief overview of what that toolkit really comprises of. So, for us, there are four big modules or four pillars, if you will. The first being around what we call strategy, but that also includes ethics, regulations and things like that. So very, very important so. The strategic pillar, there's the control pillar. This is more around governance and making sure that, beyond the data scientists, who should get involved in the governance of these AI models. We have a pillar that's focused on what we call really the responsible cases. So, testing for biases, interpretability, those kinds of measures that we do. And then the last pillar is, and last not least, because I'll say this is a pillar that often a lot of our clients and some of our internal projects as well will lack in maturity and it's more around what we call core operations. So, managing long term and really monitoring the models in AI. So, with that in mind, what I wanted to just give you guys a highlight is how holistic the toolkit really is. And what that means is that there is lots and lots of stakeholders involved in the design, the management, the control, the governance of AI. And really what that helps do is bring more people to the table to have the conversations. And also, of course, we need to educate them. They need to learn. They need to get invested, and now they share the accountability. So what I found that really had as an impact for a lot of the people, who really dove in with us and started using it, was lots more AI application ideas. So applied AI became a much broader group of people that had those interest and pushing all the way from the top. So, one of our recent clients…Of course, the generative AI buzz has helped as well…But before even that, we found that, the more people got educated and comfortable that these types of machines can be managed, the more openness there was. So just going back to that example. Right now, it's coming all the way from the CEO saying, like, we must use this technology, we must use this new machine that got developed for us. And it's really coming from the top. So again, broadening the education, broadening the accountability, is the biggest impact that I've seen when it comes to using that toolkit and bringing that holistic perspective when it comes to AI.

Brian Keng: Yeah. And that's really cool because one of the things that I've seen is that people are just too narrowly focused on interpretability or bias or fairness, whereas I think this holistic view is really kind of one of the big things that that sounds like it's really having a big impact there. Allison, maybe let's turn to you. I know that you've been working on a lot of AI projects that are having positive social impact. And one of the things we talked about before was misogyny detection and correction AI tool. Maybe you can talk a little bit more about it and some of the things that it brings over the existing work out there.

Allison Cohen: Thank you so much. And I just want to say it's a real honor to be here with all of you today. This is such an important topic and just excited to be drawing attention to some of these themes. And so yes, that's right, Brian. I'm working on a portfolio of AI for social good projects. And one of those projects is highly aligned with today's theme. It's a project called Biasly. It's a natural language processing tool. And for those who don't know, natural language processing is sort of a suite of algorithms that can be used on natural language, can understand the way that humans communicate with each other in ways that we find relevant. So, it's a natural language processing tool that can actually detect and educate users about the presence of misogyny in written text. And just for the sake of defining our terms, although I know everyone in this room knows what misogyny means, we've had to be very intentional about not just using jargon in this space, but also being very clear about what we mean when we use these terms. So, our definition of misogyny is anything that is really prejudiced, exhibits hatred or any other type of discrimination against women or anything female associated by virtue of belonging to the category of women. And I think that, you know, misogyny can sometimes be interpreted in a whole bunch of different ways. But I was recently reading Chimamanda Ngozi Adichie's book called *How to Raise A Feminist*. And she spoke about how if you're criticizing a quality in women that you wouldn't criticize in men, you're not criticizing that quality, you're criticizing the woman. And that's really where the misogyny element comes in. And the models that are on the market today have luckily given space for plenty of opportunity for us to make a meaningful impact. And that's both because we don't necessarily see a lot of tools like this, but also because the tools that have been developed in this space are problematic to say the least. And that's because you can sort of deconstruct an AI model in a number of ways. One is by looking at the data that was used to train those models. When it comes to misogyny detection, very often people use models that, well… They use data that comes from social media. And quite often, social media data is very hyper-aggressive against women. And what we know is that just as problematic as that can be, so is more subtle forms that misogyny can come in. So, a focus for us has been how do we find data sets that are more representative of the way that misogyny can be expressed, both in overt ways and in more subtle ways as well. Another element of common models in this space is that they look at keywords as signifiers for misogyny. And what we know is that misogyny is context specific. You're not going to just identify it with the use of a word. In fact, if a woman uses a certain word with another woman, that can actually be reclaimed language. It's not necessarily misogynistic. And then finally, a lot of the tools being developed, research being done, is exclusively among computer scientists as opposed to engaging multidisciplinary experts in the space. And you know, a lot of the things I've just commented on have come from domain experts that I'm working with. So we're really looking to change the narrative of how this work is being done by changing the data sets that we're using, by changing the nature of the models we're working with, and even changing how the process is being done by virtue of the people we're pulling in. We're working both with a linguist and a gender studies expert in addition to computer scientists.

Brian Keng: Very cool. And I really like this multidisciplinary aspect and I think we're going to try to come back to that. But first, something that you referenced, and Annie as well, is really these new generative models. I think everyone's probably heard of ChatGPT. There are a lot of service areas for things that maybe it could do well and things that it could do not so well. So, what's your take on this, especially as it relates to diversity, equity and inclusion and some of the topics we're here to talk about today. Allison, I'll let you start and then Annie please feel free to Jump in.

Allison Cohen: Sure. Thank you. So, in the interest of defining our terms, maybe I'll just quickly define generative AI. Generative AI is any type of AI tool that can create content, whether that's written, oral or visual. So as an example, and these are odd examples, keep in mind, but this is sort of giving you a sense of how odd this technology can be. You can essentially generate this content using only a prompt. So you give it an idea of what it is you're looking for and it can generate the type of output that that you'd like. You can type in anything from, you know, create an image of a dog eating my homework in the style of Van Gogh, to write a sick note to my teacher in the form of Shakespeare. And I don't know why those examples came to mind, I promise I'm not in high school looking to get out of class. But there's this real breadth to what these models can do, and that's because they've been trained on ridiculous amounts of data. None of that data is labeled. They just shove a stupid amount of data into models that end up having a really robust mental model of the world. You can get some insight into how comprehensive a model is by looking at the number of parameters it has. And there are billions of parameters that these models end up developing by virtue of all the data that it's learning on. And these parameters are sort of its mental model of the world. And it means that it can generalize and operate in all sorts of different contexts. And yeah, do you want to…

Annie Veillet: I'll just add. The other thing that maybe that folks don't realize is that, if there are gaps, even in this insane amount of data, it can even create its own synthetic data to kind of fill in those gaps. So it can create its own data to train itself as to what it's going to produce. So just show how complex this is, and how hard it is after to be able to explain how this machine came up with whatever content it produces.

Allison Cohen: Yeah, that's a great point. There are so many risks when it comes to this technology. One of it being you don't know where it even got some of this information from or even if it is factual. A lot of generative models have been accused of, or do hallucinate, which means coming up with random pseudo facts that can create all sorts of misinformation. But then there are risks associated with giving irresponsible answers, no citing of sources, no accountability. If you put any sort of data in there, it can be used to further train those models. So don't tell it anything you wouldn't want it sharing with billions of people. Yeah. So those are some of the limitations. Go ahead.

Annie Veillet: Absolutely…So basically what we touched on, I think what you need to be careful of is all those risks. Of course, the upside of it, I think, why it's gaining some momentum is, how fast and how powerful it is that it can create at least a first draft. And that's what we've seen a lot of organizations use the outcome of generative AI these days is, they'll use it as a first draft. But you must be aware that it's going to come out as a first draft. So it's 80% ready. We have some folks that are using it just to answer questions on their internal policies, or in their internal guidelines, or the investment guidelines, and things like that. So we'll get an answer. It's about 80% complete most of the time. And then, you, as a human, can review and validate. Because as we've mentioned we we're not 100% sure how it came up with the answer that came up. But you'll get a really great start to whatever you're trying to accomplish. If these are things that are going to be considered less risky for your organization, it's more just being efficient. It's a really interesting tool. When you get into certain use cases though, that it can make decisions for people, I think this is where the diversity and the biases and all of that becomes hyper concerning, I think, for a lot of us. So, if it's making a decision whether somebody should get hired or not, whether you should get approved or declined for a loan, whether these types of recommendations from the machine. Personally, as excited as I am about the broader use of the technology, those use cases certainly make me quite nervous.

Allison Cohen: Yeah. I think that's a very apt characterization of some of the risks. We don't know how it's making its decisions. It's not at all humble in terms of how confident it was in making its decision or giving you any sort of facts that you're asking for. But also from an equity, diversity and inclusion standpoint, there was an amazing article in the New Yorker written by Ted Chiang, I believe called *ChatGPT is a Blurry JPEG of the Web,* and this is a really helpful way of formulating, at least personally, some of my thoughts on ChatGPT. It's you're taking data that's equal parts CNN and Breitbart. And you're asking a model to come up with some synthesized view of the world. And, by definition, that data that you're training it with is noisy. It's on this side. It's on that side. It's up here, it's down there. And you have to come to some sort of middle point and it ends up being a real abstraction of the reality. And so, the nuance that exists on the web is completely wiped away. And any diversity that you are hoping for is gone. Where they converge is the area that's statistically significant in the data. So those that are overrepresented already, those that are mainstream are the voices that end up being reproduced. And then from a content developer standpoint, all of the data that's taken, no consent, no recognition, and no compensation. And the implications of that can be quite significant from an inclusion standpoint and from an equity standpoint as well.

Brian Keng: Yeah, actually, you brought up a really good point. Modern AI is really powered by this massive data, typically pulled from the Internet. And you mentioned that there are a lot of biases. The Internet is not necessarily representative of the population or of the groups that we want to include. So, this actually plays into one of the questions audience had about…Because it’s trained on this data, it introduces a lot of unconscious bias that we don't quite know why it's there, what are some of the ways that we can deal with this and make sure that these systems can generate output that is not toxic or it doesn't introduce a lot of negative things that we wouldn't want? Annie, I wonder if you want to go ahead?

Annie Veillet: OK, I can start. So, of course there are some statistical tests that you can do and you can run. And I think sometimes the misconception is that we try to remove bias. It's more that we try to inject fairness. Because there are always going to be some biases. But we need to understand what are we considering fair as an outcome for this machine. So, we talk a lot about the machine, but for us, or for me anyways, it always comes down to, what is the problem we're trying to solve with it. So, with a fair outcome, what does that mean? And then you go from there. From a data perspective, once you understand where you want to land statistically to be what we consider fair, you can use the data that you already have, and I touched on the synthetic data, so the creation of fake data, we can also compensate by creating additional synthetic data to where there may be gaps. Where we feel like we should be at a 50/50 gender type team in this organization, but we're not, and there's maybe an even smaller segment within the female group that we want to represent that we don't even have a small sample right now, let's create at least a version of it, and then hopefully it grows naturally, and doesn't need to be synthetic data with time. So, there are a few techniques like that that we can do. But it all comes down to statistics. So, if you take the time to do it, you can make sure that, maybe not on the generative side, but on the more classical AI controlled environment, you can make sure that you're testing and validating. And where there are gaps, take the time to create the synthetic data to compensate for those gaps.

Allison Cohen: Yeah. And I would also add that, with generative AI, which is all prompt based, the only insight that you have into how the tool is going to perform is by testing it and testing it on representative examples. And this is really where the democratization of AI comes in because anyone could test it. You don't need a computer science degree to be able to figure out what examples would be problematic, what examples would be interesting. And then report that back to computer scientists who can then engineer those types of findings back into the model. So, there's this really real capacity to iterate on the way that the model is working and iterate in a way that includes domain experts who don't necessarily have that insight on the computer science side. So, to some extent there, I think that generative AI could be a really interesting development in the space of responsible AI. But there hasn't been enough research precedent, best practice yet to really formalize what that process should look like. So, I'm hoping that people like you will be able to create more metrics and toolkits so that people can do this well as they're leveraging these APIs which are so easily accessible now to the population.

Annie Veillet: Challenge accepted.

Brian Keng: You brought up a good point and it comes back to this idea of interdisciplinary work that we need to do. And it shows up in your background, Allison, coming from the Munk school originally, like training in International Development then moving into AI. And so you've been doing a lot of work with teams and interdisciplinary problems in this space. Can you describe, not just having these diverse teams for the sake of having diverse teams, but how was it actually helped drive some of the outcomes that you want and what are some of the benefits that you see and think about?

Allison Cohen: I love this question. Thank you. I swear I didn't plant it myself. So when I talk about AI and the need for multidisciplinary engagement, what I often use to illustrate this point is…AI is commonly referred to as a tool and I staunchly oppose that belief. Because when you describe something as a tool, you make it sound really abstract, like very scientific, like it would exist whether humans are here or not. When it comes to human-made tools, I really think there's this continuum tools that are really removed from anthropological, sociological, psychological values and beliefs and then tools that are incredibly revealing about humanity. So, on the stripped down low dimensional side, I give you the example of a wrench. You don't necessarily need to know much about human psychology or sociology to be able to use a wrench or understand how a wrench would be used, although I'm not exactly sure on how to use a wrench. But on the other side of the spectrum, you have really complex human-made artifacts like art. And I mean as anyone who studied art would know, there's so much context that you need in order to interpret what you're looking at. And there's so much context that goes into an artist's desire to even create that in the first place. And so, on this continuum, I hope you're bearing with me, but let me know if you're not, I think applied AI projects are actually a lot closer to art in terms of being this artifact of humanity that contains psychological elements that are so informed by our economy and by our society, in terms of the people that are sitting at the table, in terms of how they're funded, in terms of what projects are important, in terms of how they collect their data, in terms of the responsible AI commitment. So, all this to say is, the tools that we're developing are not agnostic to our reality. And if we want to create tools that are high impact on the ground, we need to make sure that that reality is captured in the design and development process. And that reality can't be captured by computer scientists alone. They have to be working with people on the ground, with people coming from different disciplinary backgrounds. To give you an example, I'm working on a human trafficking detection tool. And for the first couple of years, it was developed solely with computer scientists. And then we introduced criminologists, lawyers, ethicists and survivors of human trafficking, and it completely changed the nature of what we were doing. The data we were paying attention to, how we aggregated it, how we ran it through our models, what our models are looking for and our deployment strategy. So, I think multidisciplinary is the difference between having a great project in theory and then having a meaningful project in practice.

Brian Keng: Yeah. And based on my experience working with technologists, this is something they often forget. And I think this is something we definitely need to include. And Annie, you have your own experiences. And one of the things that you mentioned before is that when you joined the team you had was of 100 or so, 14% woman. And then now it's closer to being even around 40%. And you said that this explicitly has helped the models that you make be more fair and less biased. Can you maybe talk a little bit more about how that actually has produced better results having a more diverse team?

Annie Veillet: Absolutely. And I'm gonna say it 40% and growing. We're trying to reach equality. It does make a huge difference. And I think Allison mentioned it. It's that diversity of thinking… It's been an increasing diversity, of course, in the gender, but also in the backgrounds. We have folks that on our team that have a PhD in music theory. We have people that have all kinds of backgrounds now. Most of them are obviously very passionate about technology obviously, but still…Now when we're doing those first steps of the designing and the selections of the types of data we should consider to solve the business problems…and I'm mentioning this again, for us, because we work more for organizations, if we're going to train and work on a model, it is for specific business problem. So if you take the group of people that needs to look at the business problem, it's rarely just one type of profile that will come to the best, more complete solution, both for the performance of the results, but of course to be more for fair and have considered a variety of different angles. So, if I go back to one specific example, we were working with an organization. It was a pretty simple business problem in the sense that we're trying to build a model that is going to predict or recommend where we should do the next subway station or metro station, depending on the city you're in, for the communities. And the first draft of the design had a narrow set of data and it had some pretty fundamental recommendation on which methods we should use, whether we use NLP, machine learning, etc. But then when we added additional people around the table to think through the business problem, we added so many other data sets, to your point Allison, which ones were we actually paying attention. We went more on the behavioral side of things. We went more on other sources of data that were not as important in the first consideration and the first design. So, with that diversity and the team, even the data scientist team, I'm going to call it, it improves. And if you can go above that and have an even more holistic type of team solving that problem, I love the idea of bringing like lawyers at this, whoever can help bring a point of view to that business problem, should be around the table. And that's going to help, frankly, with the performance of the machine itself, making sure it's really giving you the. Result to solve your problem in a more efficient manner.

Brian Keng: Yeah. And that's really cool. And I think that having that diverse team sounds like it really helps to develop the final solution. But I guess one of the things that we also need to focus on and I'm curious to see what you think about it. During the formative years, during education, as I mentioned a lot of technologists that I work for or with, really just thinking about the technology problem, not the different perspectives, whether they be from gender or backgrounds or cultural. So do you have any thoughts about how we can improve those perspectives, especially with these technologists who haven't really been thinking about this topic?

Allison Cohen: Yes, I was actually shocked to find out that you can graduate with a Masters in computer science without having ever taken a course in ethics, which it's so terrifying considering the scale of the deployment of these tools and just how intimate they can be in our lives. I think what's also wild to think about is that ethics is really the training on how to navigate grey areas and grey areas exist all over the AI life cycle, even when it comes to evaluation, the tradeoff between precision and recall when you're looking at F1-scores. They're everywhere. So I think gaining traction is an important area for computer scientists to be learning about, especially with all the reputational damage that can be caused when the machine learning tools don't have the intended impact in practice, and can in fact end up hurting a whole host of communities. So hopefully the importance of this is growing. But it's growing slowly because the thing that's growing faster is the economic, the insane economic returns, that some of these tools can have. You can look at open AI. They've gotten investments on the order of billions from Microsoft. I mean, there's also a definite incentive to move as fast as you can deploy as quickly as possible, break whatever you need to break, even if that's trust with local communities. So I think we would be remiss if we didn't address this very strong financial incentive that's getting people to move in a way that's not conducive to trust building or to an ethical reflection which takes a long time to do. So, an antidote to what we're focusing on at Mila, which is where I work. We have what's called TRAIL, which is an education program for our researchers. There are around 1000 affiliated academics with Mila and we're running them through this supplemental ethics program to teach them about how to incorporate some of these reflections in their work. But I think there needs to be a real structural focus on how it does all this money influence the way that the technology gets built, and the speed at which it gets built, and how irrelevant questions of ethics and trust building become because of that, which I think is a big problem.

Annie Veillet: Absolutely. And if I can build on that. For us, we hear a lot of things like privacy by design like, when you're designing, you consider the privacy from the very beginning of the process. So, similarly maybe it's ethics by design. Practicing that kind of method, I think, is really missing today in a lot of the education. We try to compensate for it in the real world and add supplemental training, but certainly something that we could start much earlier to have kind of that knee jerk reaction from the get-go. You need to think about…OK, I'm working in data…of course I'm going to use data and it's going to have some results, so privacy, ethics, there is a bunch of different aspects that you need to think about early during the design phase. That's one part of it. The other thing that we've been doing in that supplemental training in the office for us, for our technical folks, but also our business folks, is… we called it Agile BXT, but really what this is…At the design stage, we bring a bunch of different stakeholders…It kind of goes with your multidisciplinary angle… We try to say like you're coming in to represent like the techs, like the CTO's of the world, you're coming in to represent the CFO…so trying to really again have a multitude of perspectives and we build the designs like that. So it's no longer just three mathematicians going in a room trying to design the most performant machine. It's really, OK, multidisciplinary, we're all looking at a different angle and then with the results of that you can go apply them afterwards. It's another type of training that we've been trying to complement what we've seen in the education system. Of course, I'll do a shout out to, if we can find ways to encourage even more women to go into the more technical fields, I think that will help as well. And indirectly, pure training. I think those are two ways of complementing or augmenting what's already taught.

Brian Keng: Yeah, that's great. And this relates to a couple of questions that people have submitted. One is about encouraging more of this collaboration across stakeholders is maybe something to help remedy this problem. But on top of that, maybe I'll add… another question they're asking is…Is regulation some of the answer to this? And I'd love to get your opinions on. What do you think?

Allison Cohen: So I think. Yes. As we're seeing there's such financial incentive for people to move as fast as possible, so unless you have some sort of regulation, at least mandating a level of ethics or compliance with established regulatory procedures or standards, you're not going to have the desire to staff teams more comprehensively, especially because it really does slow the process down to have divergent viewpoints. I've been developing this bias detection tool for the last two years and it's quite likely it will continue to take at least another year, if we're going to make it, a functioning prototype. And I'm sure that there's companies want to do the best that they can get away with. I don't think they want to do the best they can possibly do. So in order for us to see the types of products that are going to truly benefit our societies, I think that's going to come from regulation insisting on that kind of process.

Annie Veillet: Yeah, I concur. Yeah, I think for those who are a bit in the world, like there's a Bill C27 coming, we think 2025, some groups are trying to pressure to speed that up. We certainly appreciate... I think for anybody working in this field, the last thing we would want is for folks to try to go too fast and cause accidents, if I could call that, in this world, and then have everybody be overly scared of, I think, this fabulous technology really. So I think yes, regulations, because it's going to keep everybody safe. And there are ways to use these techniques to be in a safe way, in a fair way, if we take that time. But it does cost a little bit more money to do it right. So when we go with the organizations who say, like to be compliant, you need to do XYZ, typically the budgets get freed-up. If it's just because it's the right thing to do, there's a little bit, a lot more, I would say, friction to go get the budgets required to do that more holistic view. Because it's not only…we've talked a lot about design recently, but it's also the governance after, and the controls, and the maintenance at the end. Because models can shift with time, so it's not only how did you design it, but are you monitoring to make sure it's still doing what it's meant to be doing to solve that business problem. So multitudes of reasons why I also would appreciate and applaud any acceleration in some of these bills being passed to add regulations.

Brian Keng: Great. So, we have a bit of time. I'm going to turn to some of the questions from the audience. This is one that came up. Do you think current models can detect dog whistles, misogyny, racism and transphobia, etc, given that they're so context specific and designed to be hidden? So this is, I guess, maybe more adversarial from the human point of view.

Allison Cohen: This is a great question. In case anyone…I wouldn't have known what a dog whistle was unless I was working on this project, so I think the question is sort of asking what about the language that's meant to go undetected, when you know people are communicating with some in-group and trying to signify whether it's hatred, bigotry, misogyny. This is a real problem and it's a real limitation of our tool. I think this speaks to all of the possible oversights that a tool like the one I'm developing could have. When it comes to ethical AI, of course, a big part of it is the tool itself and how it functions. But you also have to be very clear in the deployment of your tool, what its limitations are, making sure that users are using it as intended. The goal with the tool that we're developing now is to help people see what is obvious to a gender studies or linguistics expert, not necessarily to someone who specializes in dog whistles. And so, you really want to make sure that the impact that you can have isn't overshadowed by all of its limitations. But you are humble in the application of your tool, so it's only used in the way that that you're intending it and the way that you've trained it to operate. But no, we will not, unfortunately, at this stage in the game, we will not be able to identify encoded language. Because, Annie as you mentioned, these things change over time. If our tool could detect dog whistles, those would no longer be the dog whistles, and there'd be new dog whistles. And so, then you have to consider the model direct question. So yeah, I think it's important, especially as you're using AI tools to really think about the scope in which they should be responsibly applied and the scope in which they shouldn't be.

Brian Keng: Great. Maybe the next one here is. So we talked about how a lot of these modern models are built about noisy data, usually scraped from the web. Is there an effort to try to filter down the data set so that you can really get it to a place where it's representative, fair, etc. and then train the model on that? Is that even feasible from the point of view of getting enough data, clean data, to do that?

Annie Veillet: I would say absolutely and that's why I mentioned at the design stage, there’s different techniques. So, if you're going to train a model that's going to make a decision, use the example of if you're going to get approved or declined for a loan, you probably want a static set of data that you've really cleansed, that you've prepared, that everybody feels is a fair representation of your audience basically for whatever the machines going to decide. So you spend that time preparing that set of data to train that model and then you don't allow the model to continue learning on the human noise that can get created by new decisions being made by folks on the teams, et cetera. So, you can absolutely have some AI models in a controlled setting. And for those kind of more sensitive decisions, it is still probably the safest way to do it. There are ways to put guardrails on self-learning machines, but you are more at risk for the additional noise. And noise comes from the humans creating more data. And we all know that humans are very biased. So if you allow that, you have to understand that you are most likely then at risk of including additional biased data, if I could call it that.

Allison Cohen: Yeah, I would completely agree. And I would add that another way that human bias gets implanted into the data is with the data labeling. So, when you're doing data labeling, also called annotation, you hire people to pretty much give the data that you're interested in a label of a certain type. And a lot of AI projects are actually using really tricky software to do this annotation. For example, a lot of computer scientists are using platforms like Amazon's Mechanical Turk, where you have annotators that are not being properly paid, that are not being incentivized to provide the most high quality label to your data, that are incentivized to move as quickly as possible to get paid as much as possible. And the labor conditions are obviously just very questionable. So, I think, even though labeled data is considered to be a bit of a gold standard, you have to look under the hood of what that gold standard labeled data set is, because in some cases it can be data labelers who are providing annotation on English language documents, for example, and English may not be their first language, or they don't care to read an entire document for $0.10 an annotation. So really getting into the politics around annotation and the biases that can introduce is also a really important aspect of understanding the quality of the data set.

Brian Keng: OK. So we only have a couple of minutes left, but I thought this question was really good. Is there a place we think AI shouldn't venture into? You know, we talked about AI art and stealing some of the IP from humans, maybe very quickly, both of you, if you could give your viewpoint on it.

Annie Veillet: For me. if well governed, if done well, and if done in a way that we are willing to consider fairness all the way from the design to the monitoring, I think this technology can enhance and really improve a lot of our lives frankly. So there's not an absolute, don't go there and be collaborating. And for me, I'll add one thing though. This is always a collaboration between the human and the machine, right? So it's never a situation, even for the most mundane things, I don't particularly like to say, like let it all be handled by machines and the humans will do something else. So, in a safe environment, I'm up for at least consideration and with that diverse team. If there's really a risk that we feel we can't handle, that's when I would pass on it, but before that, I say let's look at it and see if we can have a safe way to do it.

Allison Cohen: Yeah, I think that that's the view of many people. There's so much potential here and no one wants to preemptively say, Oh no, AI should never go there…because maybe the impact can be significant and meaningful. But I think we need to focus on who has a say in determining whether or not the technology should be deployed there and the focus should be the communities it's affecting.

Sarah Kaplan: That conversation stimulated a lot of ideas for me. I was particularly drawn to the idea that you need interdisciplinary teams, including social scientists and not just computer scientists, to design equitable AI. And, you need to design with the risk of inequalities in mind at all times. At GATE, we've been thinking about these issues for awhile now and we'll post a couple of resources for you that should take the conversation further.

Thank you for listening to this special edition GATE Audio production podcast on "Designing for Everyone." If you haven't listened to them already, I hope you will check out the other 6 episodes in this limited-edition series and other GATE Audio podcasts, including our signature podcast, BUSTED, where we bust common myths about gender and other forms of inequality. Just search for "Institute for Gender and the Economy" where you get your podcasts. Of course, you can help us get the word out by liking and following the podcast and telling your friends. We are nowhere without our community of listeners. If you want to keep learning, head to our website at GenderAnalytics.Org where you can discover our online course offerings and much more.

This podcast was produced by me, Sarah Kaplan, and edited by Ian Gormley. We are grateful for support from the Rotman School's TD Management and Data Analytics Lab who co-hosted the Gender Analytics: Possibilities conference with GATE.

See you next time!