SIGecom Winter Meeting 2022 Highlights

EMILY DIANA
University of Pennsylvania
and
MINGZI NIU
Rice University
and
GEORGY NOAROV
University of Pennsylvania

Emily Diana *is a rising fifth year Ph.D. student in Statistics and Data Science at the Wharton School, University of Pennsylvania, where she is advised by Michael Kearns and Aaron Roth. Her research focuses on the intersection of ethical algorithm design and socially aware machine learning, and she is honored to have been recognized as both a Rising Star in EECS by MIT and a Future Leader in Data Science by the University of Michigan. Before Penn, she received a B.A. in Applied Mathematics from Yale and an M.S. in Statistics from Stanford, and she spent two years as a software developer at Lawrence Livermore National Laboratory.*

Mingzi Niu *is a rising fifth year Ph.D. student in Economics at Rice University, where she is advised by Mallesh Pai and Hülya Eraslan. Her research interest are primarily in microeconomic theory, with a focus on mechanism design, information theory and behavioral economics. Before Rice, she received a B.A. in Finance and Banking and a B.S. in Mathematics and Statistics at Peking University, and a M.A. in Economics at Duke University.*

Georgy Noarov *is a rising third year PhD student in Computer and Information Science at the University of Pennsylvania, advised by Michael Kearns and Aaron Roth. Previously, he graduated from Princeton University with a B.A. in Mathematics. His research interests span across the fields of uncertainty quantification, online learning, fairness in machine learning, and algorithmic game theory.*

The Second Annual ACM SIGecom Winter Meeting took place virtually on February 23, 2022. Organized by Mallesh Pai and Aaron Roth, it brought together researchers from economics and computation and adjacent communities to focus on the the topic of Fairness (broadly construed). The 2022 Winter Meeting featured tutorials and invited speakers spanning many disciplines, as well as a fireside chat and other social activities. We present some highlights from the event, including a recap of the fireside chat with Cynthia Dwork and Sendhil Mullainathan, and interviews we conducted with invited speakers.

Editors' note: *Due to an error from the editorial team, an interview with Juba Ziani, who gave a tutorial at the 2022 Winter Meeting, was not included in this issue of the SIGecom Exchanges. The interview with Juba Ziani can be found in the Winter 2022 issue.*

Recap of the Fireside Chat with Cynthia Dwork and Sendhil Mullainathan

One of the highlights of the 2022 Winter Meeting was the Fireside Chat, a 30-minute long Q&A session with Cynthia Dwork (Harvard) and Sendhil Mullainathan (UChicago). Both panelists are renowned scholars and authors of seminal research on Fairness in ML. The Fireside Chat was an exciting part of this workshop, giving food for thought to both young researchers and seasoned scientists wishing to enter the field of algorithmic fairness. Here are edited excerpts from several questions that the panelists shared their wisdom on.

Fair ML theory experts often face the criticism that mathematical fairness research is reductionist and too narrowly focused. As a result, some may argue that it brushes aside real-world structural issues of injustice that do not have straightforward technical solutions. Is this a fair criticism?

Mullainathan. Mathematical fairness researchers think like philosophers. They seek to design a language and framework for discussing and engaging with fairness issues. Mathematics simply extends this philosophical mindset by allowing us to make our statements even more formal and precise. Both theorists and philosophers approach fairness issues broadly, just like humanists would. They search the space of possible definitions and notions of fairness and investigate their interrelationships. In parallel, they constantly perform reality checks on these definitions and look for missing pieces that could be added to the theory. Indeed, any single paper on fairness will typically only look at a specific and narrow aspect of fairness — but together, these papers form a broad and diverse body of research, whose goals and breadth are fundamentally in line with what humanists attempt to do.

Dwork. Mathematical fairness research is essential for the field's future success. This is analogous to how mathematics has revolutionized the field of cryptography: mathematically formal cryptographic protocols and schemes have been instrumental in enabling engineers to build powerful and scalable code for complex cryptosystems. This process of making cryptography rigorous has been taking place since World War II, and has helped us formally reason about crucial questions such as: What exact security guarantees are we trying to achieve? How powerful is the adversary we are defending against? Clearly, modern-day cryptographic software would not have been possible without first attaining this high level of mathematical clarity and precision. Similarly, the field of algorithmic fairness is currently going through the cycle of proposing new mathematical definitions, augmenting them, and proposing new ones. In this manner, we are following a clear path of progress providing an indispensable foundation for concrete, in particular software-based, future fairness solutions.

Over the last few decades, many predictive models have become central ingredients of automated decision-making tools used by governments and businesses. When deployed in areas such as hiring, lending and law enforcement, the decisions made by these models directly impact people's lives, potentially in negative ways. For instance, neural network-based facial recognition tools are widely used in law enforcement to identify

criminals, but they are known to be prone to having baked-in racial biases. Can a researcher developing an ML model predict and forestall any future long-term fairness-related risks that the model may pose once deployed?

Dwork. The only, and indispensable, way of identifying and preventing future fairness issues with a model is to take time before deploying it and speak seriously with lots and lots of different groups. This will help you see how what you are doing is received and whether it is perceived as appropriate or inappropriate. This question also directly links to the issue of responsibly scaling AI solutions, which is something that tech companies — including behemoths such as Meta and Google — have been increasingly grappling with.

Mullainathan. Our aim should not be to develop extreme degrees of foresight into such future issues. Rather, we should identify fairness flaws in AI models by exercising vigilance. Often, influential AI models can in a matter of 5-10 years become impactful beyond all our initial expectations — and just as they turn out to be unexpectedly powerful, they may become dangerous in various unexpected ways. As a result, we cannot hope to reliably predict fairness-related fallouts — but we can continually monitor the situation to identify any emergent fairness risks.

Can you identify a "Greatest Hits" list of Fair ML papers and books that all researchers entering the field should study?

Dwork. The paper *Fairness through Awareness* [Dwork et al. 2012] initiated the study of fairness in machine learning. Among other things, it articulates and elaborates on the difference between individual and group notions of fairness. *Inherent Trade-offs in the Fair Determination of Risk Scores* [Kleinberg et al. 2016] is a seminal paper that demonstrated a fundamental conflict between several very natural definitions of group fairness.

Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness [Kearns et al. 2018] and *Multicalibration: Calibration for the Computationally-identifiable Masses* [Hébert-Johnson et al. 2018] contemporaneously introduced multigroup fairness: a setting where fairness guarantees are given for each group in a potentially complex (e.g. large and intersecting) family of population groups. Multigroup fairness can be viewed as providing a bridge between individual and group notions of fairness.

Mullainathan. The field of algorithmic fairness is still in its budding stage, so we have ample opportunity to contribute to the literature by coming up with novel models of real-world phenomena we care about. By contrast, in many well-established research areas a lot of modern-day research is a further elaboration of existing models. An excellent general-audience book illustrating how recent fairness research connects with real-world issues and phenomena is *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* [Kearns and Roth 2019].

Interview with Annie Liang

Annie Liang is an Assistant Professor of Economics and Karr Family Assistant Professor of Computer Science at Northwestern University. Her research is in economic theory, and the application of machine learning methods for model building and evaluation.

As an invited speaker at the SIGecom 2022 Winter Meeting, Dr. Liang gave a presentation on her recent work "Algorithmic Design: Fairness vs Accuracy". This paper is joint work with Jay Lu and Xiaosheng Mu.

Dr. Liang's talk described an elegant framework addressing the important and delicate issue of balancing accuracy and fairness in automated decision-making. For concreteness, imagine an automated hiring process where a decision-making algorithm receives features of candidates coming from two different population groups, and outputs a binary hiring decision for each candidate. We (the designer) can control which decision-making algorithm is used, as well as regulate what information the algorithm can access about the candidates. The central object of study is the accuracy-fairness Pareto frontier, which characterizes all "optimally fair" ways for us to trade off the algorithm's performance (i.e. its error rates) on each of the two population groups. Many natural notions of fairness are permitted, including the egalitarian (the group errors must be similar), the Rawlsian (both group errors must be small), and the utilitarian (the overall population error must be small).

Even though it is simple, this framework is surprisingly rich and yields plenty of rigorous qualitative insights on the accuracy-fairness trade-off — in particular, on the implications of the algorithm using or ignoring the candidates' group identities or group identity correlates. This helps cast new light on some hotly debated real-world topics such as the Ban the Box movement, and the ban on using test scores for admissions purposes implemented by some US colleges.

In our post-meeting interview, Dr. Liang told us more about this exciting research project, and also spoke about her academic trajectory and her perspective on the growth and development of the econ/CS community.

I would like to start by asking you about your academic path so far. What brought you to the econ/CS intersection?

I've had an outsider's respect for CS since undergrad at MIT: I was myself an economics and math major, but computer science was very big there, and I absorbed this idea that computer science was really cool. I didn't personally get interested in computer science until I was in graduate school in economics. The initial hook for me was different definitions of complexity, but my interests quickly grew from there. And it was a good time to be thinking about CS, since economists were starting to become aware of and get excited about machine learning. Several of the faculty organized reading groups with students to learn about new ideas in CS. And later, I was fortunate enough to do a postdoc at Microsoft Research, where I was exposed still further to computer scientists working at the econ/CS intersection.

What did the econ/CS area look like from your perspective back then? And which directions are you excited about right now?

When I was in grad school, CS/econ was still a niche area — that has changed quite a bit in the last couple of years, which I am really happy about.

I think there are several interesting directions at present. For example, there's the growing area of econometrics and machine learning (related to causal inference), and machine learning is increasingly used as a new tool in empirical economics. I've personally been most involved in the intersection between machine learning and economic theory.

In principle, there's a conflict between the machine-learning, or black-box, way of doing prediction and the way that economic theorists think about model building. Economic models tend to be interpretable theories that offer some narrative or explanation about the underlying behavior, while black box machine learning models are often complicated objects, where it isn't clear why the black box is predicting what it is. But I've always thought that there were potential complementarities between these two methods, and a lot of my recent work has been about how we can use black boxes to better evaluate or improve on economic models.

In the other direction, I think that economics has a lot to add to computer science as well. In the last decade or so, computer science has taken a turn away from developing algorithms with clear, well defined criteria in mind — such as predictive accuracy – to considering these algorithms within a larger social context. Economists can definitely contribute a lot here, because economics has had a long history of developing frameworks and tools for reasoning about social welfare.

What specific benefits do you think economic modeling can contribute to future fairness research?

One of the things I am most excited about regarding the paper that I presented at this workshop is that we were able to import an economics perspective on welfare and preferences in these settings.

Much of the literature in CS on algorithmic fairness literature has focused on a specific optimization criterion—for example, showing how to optimize for efficiency, subject to constraints such as equalized error rates. In the paper that I presented, we defined a broad class of different preferences that the designer might have, varying across many different ways of trading off between fairness and accuracy: from utilitarian preferences to pure egalitarian preferences. Many insights in this paper hold without needing to specify exactly what the objective function is. There are even certain policy recommendations that hold uniformly across this diverse class of fairness-accuracy preferences.

Broadly generalizing, I think there is a tendency in computer science to want to provide a *solution* to the problem at hand. This is a bit different from the way that economists approach the problems that emerge in social science, where the goal is sometimes not to provide a solution, but rather to figure out how to think about the situation and the inherent trade-offs. There's clearly value to both approaches.

The paper you presented at the Winter Meeting offers nice geometric insight into the nature of this fairness-accuracy Pareto frontier. At the

same time, this is enabled through the setting being two-dimensional: in particular, the paper mostly deals with two population groups, and the case where you can only undertake binary actions (such as making an accept/reject decision) for each data point. How hard do you think it would be to extend these results to more groups or more actions while keeping the geometric insights intact?

Extending this theory to more groups is probably not difficult. One would need to decide on how to generalize the fairness notion: should we be comparing groups individually, or relative to some average, or looking at the worst-off and best-off groups? But ultimately, this choice probably will not have a qualitative impact on the results.

Regarding more than two actions: the full design case in the first half of the paper extends readily. It's once we bring in the information design problem that we actually start using the binary nature of the actions.

How did this project begin? Did it stem from you and your collaborators thinking about various fairness problems out there, or perhaps from a concrete mathematical problem?

I've been aware of the algorithmic fairness literature in computer science for quite a while, and already during my postdoc people were very excited about it. So I have been following and admiring this literature, and I definitely wanted to write a paper on this topic. My collaborators and I were especially intrigued by the trade-off between fairness and accuracy rates: it seemed evident that such a trade-off might occur when groups have different distributions, and we wanted to know if we could say something about it. All three of us have experience working on papers in information economics, and we naturally were also curious about how the information fed into the algorithm affected the nature of this trade-off. So we began by sketching out some formal models to think through these questions, and went from there.

How fast do you anticipate the econ/CS and fairness areas will be developing in the near future? And what does this mean from the perspective of young researchers in these areas who are planning to enter the job market?

There are many signs indicating that this is an emerging and rapidly developing area. In the last couple of years, graduate students in economics have been increasingly going on the job market with econ/CS papers and getting great jobs. I've also been noticing more and more job postings explicitly looking for somebody in this intersection.

I know you have participated in organizing several great workshops focused on game theory, and on the social impact of machine learning. Naturally, these events bring together researchers with unique perspectives on fairness and econ/CS. Could you speak about your experience organizing the workshops and bringing together all these different speakers, and any takeaways?

As I mentioned, this intersection is growing very rapidly — but I find especially interesting that it is growing in many different directions simultaneously. In general, I think one shouldn't picture the econ/CS intersection as economists (as a monolith) interacting with computer scientists (as a monolith): there are many exchanges going on here, and many different subcommunities involved. For example, my initial exposure to computer scientists was through the algorithmic game theory community, and I only realized after a while that there was a separate machine learning community, with a different (but overlapping) group of people and set of conferences. And in the same way, economic theorists are not the same as econometricians who are not the same as empirical economists, although each of these groups has recently been shaped in some way by computer science. So the most interesting takeaway for me so far has been the vast diversity of these synergies between the two fields.

It will be interesting to see how this evolves. Will there ultimately be a CS-Economics field housing all these different people? Or will each area within economics and within CS be influenced by this interaction in a different way?

As a final question, can you say a bit about your hobbies?

I've been a learner of Russian since grad school. At some point, I may decide it's good enough and move on to something else, but right now I am still really enjoying continuing to improve my understanding of the language.

Interview with Ariel Procaccia

Ariel Procaccia joined the Winter Meeting as an invited speaker to discuss his paper "Fair Algorithms for Selecting Citizens' Assemblies." Dr. Procaccia is Gordon McKay Professor of Computer Science at Harvard University and works on problems related to artificial intelligence, algorithms, economics, and society, and he is especially excited about projects that involve both interesting theory and direct applications. Most recently, his sortition algorithm and online framework at Panelot.org has been rapidly adopted by government agencies for their selection processes to form citizens' assemblies.

Dr. Procaccia was gracious enough to agree to speak with us about fair division, Panelot, and several tidbits of his experience as an academic.

Algorithmic fairness has exploded in the past few years. Where do you see the field going? Similarly, what do you think are the most important open problems and areas for future research in the field right now?

Let me mention one direction that I think is important. The study of fairness in machine learning has developed almost independently from fair division, an area that dates back to the 1940s and has very similar goals: to define rigorous notions of fairness and devise methods for achieving them. Not surprisingly, notions developed in fair division can be applied to fair machine learning. For example, the classic notion of envy-freeness can be used to design fair classifiers: the utility of each individual for their own (possibly random) outcome should be at least as high as their utility for any other individual's outcome (this makes sense when utilities are heterogenous). Going forward, I believe that ideas from fair division, and, more generally, from normative economics, have a much bigger role to play in fair machine learning.

Can you tell us a bit about your experience with Panelot and working with government officials? For example, how did you get interested in algorithms for political fairness purposes? Was it hard to get your algorithm publicized and launched?

I've always been excited about the intersection of computer science and democracy. I got interested in sortition – random selection of representatives – specifically, when my Ph.D. student Paul Gölz recommended to me an amazing book, "Against Elections" by David Van Reybrouck. In 2019 I wrote an opinion piece about sortition, which led to conversations with practitioners. Eventually Paul and I were invited to a demonstration of an algorithm for selecting citizens' assemblies, which was developed and presented by Brett Hennig of the UK-based Sortition Foundation. This was the beginning of a wonderful collaboration with the Sortition Foundation, which later facilitated the deployment of our own algorithm and its adoption by other organizations.

How did you handle the code development and professional software production for Panelot?

Our selection algorithm was coded up by Paul, and in the initial deployment we simply plugged his code into the open-source interface created by Brett. The website Panelot.org, which makes the selection algorithm more easily accessible, was mainly created by Gili Rusak, who was a master's student at Stanford at the time and will start her Ph.D. at Harvard in the fall. Other contributors (who also played key roles in designing the algorithm) include Bailey Flanigan and Anupam Gupta. To summarize, code development and software production were all done on a pro bono basis by our research group, and the code is open source. (That said, there are some expenses, including the design of a professional logo and, more significantly, running Panelot on AWS.)

Do you have any advice for young researchers?

My number one advice for young researchers is "frequently say no." Academia has an unusual workflow in that one is asked to do many things (reviews, program committees, talks, department service, etc.) by many people who are not aware of each other's requests. This issue is especially acute for young faculty members, who typically say "yes" to almost everything and end up being inundated with tasks they can't complete. Be judicious about what you agree to do.

What do you enjoy doing outside of research?

I have three kids (13, 8 and 3) so between family and work I don't have a lot of free time. But one thing I still greatly enjoy is playing video games. Currently I'm perhaps 60-70 hours into Elden Ring and, disturbingly, the game claims my progress is 20%, so I expect to finish it around 2025.

Interview with Hoda Heidari

For the last talk of the Winter Meeting, we had the pleasure of listening to Hoda Heidari discuss her recent paper with Jon Kleinberg, "Allocating Opportunities in a Dynamic Model of Intergenerational Mobility."

Dr. Heidari is an Assistant Professor at Carnegie Mellon University with joint appointments in the Machine Learning Department and the Institute for Software Research. She is broadly interested in societal aspects of artificial intelligence and machine learning and, in particular, algorithmic fairness and accountability.

Dr. Heidari was kind enough to give us an inside perspective on her paper and share her broader experiences as an academic in this growing field.

I see that you recently joined Carnegie Mellon as a faculty member. How has your transition been?

It's been great. My job responsibilities as a faculty member are different compared to when I was a doctoral student or postdoctoral fellow, so definitely the volume and diversity of responsibilities amps up substantially, but I also have the privilege of advising students and teaching classes that I enjoy. So overall, I have more on my plate, but at the same time more autonomy and opportunities to push my research agenda forward and contribute to training the next generation of researchers in my field.

I see you are teaching a class "Machine Learning, Ethics, and Society." It sounds exciting to be teaching a course on very new material – is it difficult to have it be comprehensive and fit together as one unit?

It definitely is – as you mentioned, the material is very new and the research community as a whole is still trying to figure out its path and purpose. Currently, my approach is to offer a sample of the existing research landscape. I hope that at the end of the semester the students see a common thread, but I don't offer the topics as sequentially related to one another. That's something reflective of where the research is, and I think it is, in a sense, liberating. There are not that many standard topics and methods you feel obliged to cover, so you get to shape the syllabus to teach students how to critically evaluate new situations and problems they may face in the future.

And have you been finding that there are specific things that the students get really excited about?

It's amazing how engaged the students are in the class discussions. I make sure to have multiple open-ended discussions because this is not a topic you can cover through a one-sided lecture in which you tell them what is the correct or right way of looking at the problem. It's important to stimulate students' own ways of reasoning about a new scenario. One area that my students often passionately express their thoughts and experiences on is the issue of fairness, and one common theme in

their comments is about the limited and narrow nature of existing definitions of fairness. Their concern is justified as much of the mathematical modeling that has been done around fairness is really about a very narrow definition of parity in predictive outcomes, and they are only valid for very specifically defined decisions. They do not capture a whole lot of other important factors including procedural and social justice considerations. What was the process by which the decision-making system came to be? What is the institution governing system? What are the checks and balances around it? What does fairness even mean in the specific context of decision-making? These are reflected in students' questions asking why we are focusing on the specific definitions of fairness and the research community hasn't moved on to much broader notions. I think the students always rightly point out that the existing formalism somehow feels very limiting and I think they're absolutely right. One way in which I think we can address some of these concerns is through effective engagement with stakeholders and impacted communities. So I make sure to have a module in my course that brings in community-oriented approaches and perspectives.

One question that I had about your paper modeling affirmative action policies is whether there was interest in implementing these decision-making procedures and trying to do any behavioral studies. Is there an application coming up?

I am not sure – with a stylized model of the type we proposed, the point is not necessarily to claim that the model is sufficiently realistic to warrant real-world applications. Rather the purpose is to strip away all sorts of nuances and layers of complication from the real world, so you can rigorously analyze a very stylized, hypothetical world instead. There were two key points that we were trying to make with our analysis. One relates to the discussions around the temporary versus perpetual nature of affirmative action. Usually affirmative action-type policies are cast as temporary measures to level the playing field, but it is not clear what that means and what would be a good way of phasing it out. The second point relates to "fairness interventions" – for instance, enforcing statistical parity is not exactly the same as affirmative action, but it clearly is conceptually similar. So we should carefully consider the dynamic consequences of enforcing such fairness constraints. It's not just that we are employing this intervention today and we are done tomorrow, but rather we should think about what the impact of it is on the underlying population and how we should update the model moving forward.

Finally I should emphasize that for a highly charged topic like affirmative action, implementation is usually something that is impacted by many political considerations, and we definitely don't think our work on its own is sufficient to inform such decisions.

Yeah, that makes sense. I like that you were able to frame the problem in a way where you weren't explicitly weighing the moral implications of affirmative action but rather taking a long-term utilitarian perspective. I'm curious what pushback you got in this paper.

If you look at economic models of affirmative action, one aspect that's usually accounted for is the strategic component of agents' behaviors. A typical model

would assume that parents have a certain amount of endowment, and they decide what portion of it to invest in their offspring. We basically ignored strategic considerations and instead focused on the temporal aspect, which was something that was absent from prior work. So the absence of strategic modeling was one valid criticism. The other more conceptual criticism was that some readers had a hard time distinguishing between *socioeconomic* affirmative action and affirmative action based on *demographic* characteristics such as race or gender, and the moral implications are obviously very different. We always try to be very clear about that distinction when discussing our work.

Your primary location is in computer science, but you also work in the intersection of computer science and economics. Do you find that there are a lot of challenges translating between the two communities? Do you think that they complement each other well?

Well, having thought for a while about the ethical considerations around machine learning and AI (which are topics in humanities and social sciences), I have realized that the synergy between economics and computer science is already great. There is a common language that people in both fields speak. Game theory, for example, is one such tool used by both computer scientists and economists. We may be more interested in algorithmic problems and they may be interested more in the modeling component, but at the end of the day, we all do math. Now that my work has shifted more towards ethical considerations such as fairness and explainability, I have started collaborating with scholars in philosophy, law, sociology and so on, and we are still in the process of forming that common language. There is already a tradition of computer scientists and economists working together, which has been going on for almost two decades now, whereas the collaborations between computer scientists and scholars in social sciences and humanities is very new. So we are currently at a formative stage – definitely uncertain, but at the same time, immensely exciting.

What is your process like for coming up with problems?

I don't think I have a very systematic way of coming up with problems. Reading and talking to people, that's my main source of inspiration for choosing research problems. I find that interdisciplinary exchanges – having conversations with colleagues from different fields, people who have different views or experiences on a topic – are fantastic sources of inspiration for research, so I try to maintain those conversations.

Finally, who is one person who has been quite impactful on your career, not counting your doctoral advisors?

Jon Kleinberg has always been a source of inspiration and my academic role model, and I have been lucky enough to collaborate with him closely over the past few years. When I was a doctoral student at Penn I took a microeconomic course taught by George Mailath. He was one of the best teachers I have seen in my life and definitely what I aspire to emulate as a teacher myself ... although I realize it will likely take twenty to thirty years of research and practice to become a teacher of that caliber!

Interview with Ashesh Rambachan

Ashesh Rambachan presented his job market paper, "Identifying Prediction Mistakes in Observational Data" in the Winter Meeting. Dr. Rambachan completed his Ph.D. in economics at Harvard this May, will be a Postdoctoral Researcher at Microsoft Research in 2022–2023, and will join MIT's economics department as an Assistant Professor in 2023. His job market paper investigates how to identify systematic errors in human decision makers' prediction-based decisions when their preferences and private information are unknown to the researcher.

After the talk, Dr. Rambachan generously shared with us about his research and graduate study experience.

Your paper "An Economic Perspective on Algorithm Fairness" with Kleinberg, Ludwig and Mullainathan briefly mentions two opposite forces at work in automated decision-making: the algorithm may simply reflect or correct the bias in data. Could you say more about the interaction between bias and machine learning?

People have reasonable intuition that if you train on historical data that is generated by some discriminatory process to then estimate some machine-learning-based model, the model's predictions and the resulting decisions may be discriminatory as well. So we try to point out the ways in which that intuition is not quite right. After we estimated some model to predict outcomes as a function of observable features, we ultimately have control over the decision rule that translate these predictions into decisions. We can implement a decision rule based on that prediction function to equalize decisions, whatever our notion of fairness is. This is the point we want to make in my paper with Jon, Jens and Sendhil.

In the "bias in, bias out" work, we try to think about under what conditions the discriminatory data generating process leads to a discriminatory prediction function. To give you an example, from the pretrial release system, we ultimately want to predict whether or not a defendant will fail to appear in courts, but we only observe that outcome if a judge decides to release the defendant historically. One model of a discriminatory data generating process is that the judge is taste-discriminating against minority defendants, meaning the judge sets a higher threshold for the minority group when forming predictions about the probability of release. However, a higher threshold for release for the minority group implies that minorities that are released are positively selected on unobservables or the judge's private information. Then among the released defendants, minority defendants actually show lower risk. In this example, discrimination in the data generating process will actually yield a prediction function that is more favorable to the minority group.

It seems to me that the majority of the literature looks at issues such as potential biases, or under-representation for minority groups, but sidesteps the ethical aspects about whether or not depending important

life decisions solely on measurable data is the right thing to do in the first place. What is your take on this?

I think that is the heart of the problem in the application of data-driven decision-making tools in social policy domains. The data we have to train these models were generated by human decision makers. As a result, economics values and emphasizes two key features of these settings. First, heterogeneity: there may be important differences across individuals in how they make decisions. How does that affect the data, and in turn affect the models we train on that data? The second is unobservables: individuals may observe extra information that is recorded, and so we have to account for them and how they affect the training data. One thing I want to emphasize is that there may be certain conditions under which we can actually formally test whether human decision makers actually have valuable private information available to them. That is what I do in my job market paper, and I think it is an important diagnostic for users of machine learning tools in high-stake policy settings – first ask could there be private information/unobservables that could explain human decision makers' choices. If so, this may serve as a strong argument for not using automated decisions. Or we could ask what is that private information, could we in principle go and collect such information.

One key result in your job market paper is a sharp partial identification of correct beliefs under expected utility maximization. If there's no belief in this range that rationalizes observed choices, then the paper interprets it as "incorrect beliefs". Alternatively, the gap between observations and model implications may arise because the expected utility theory itself is not descriptively accurate. Could it be that people are making the right decisions, and that we researchers try to use some oversimplified model to fit human behaviors?

First, expected utility theory is the natural benchmark that every empirical researcher in economics is going to reach to model decision-making under uncertainty. So understanding what we can and cannot say about beliefs under expected utility framework is a valuable exercise. Second, I also think of it as a normative restriction on behaviors. For example, it is reasonable to say that the policy makers in the pretrial release setting would like judges to act as if they were expected utility maximizers. Now you could question whether we actually want decision makers to be expected utility maximizers in the first place as opposed to some other decision-theoretic criterion. That objection highlights that when we deploy machine learning tools, it becomes even more important for policymakers to be extraordinarily explicit about (a) the objective function and (b) how exactly they want that objective function to be maximized.

Could you share with us how you came up with ideas for your job market paper?

I came up with this idea because I've been spending a lot of time thinking about the use of data-driven tools and the related econometric issues. One thing I kept coming back to is a very simple question: why exactly would a policy maker want to replace a decision maker with an algorithm? What are forces that are on the table?

After conversations with a lot of people, I believe there are three key forces at play in these settings. One is that the policy maker may worry that decision makers mispredict based on observed features. Second they may think that decision makers have an objective function that is misaligned with the policy maker's objective function. Third, decision makers may have private information. So I wanted to think about how I can actually use data to test whether these forces exist, what the magnitude of these forces is and how understanding these forces could in turn inform the design of algorithmic tools. So that's where this paper came out.

When writing your job market paper, did you come across any major obstacles or changes in direction?'

I would say for me the biggest turning point was thinking about an actual application. Only once I had been thinking through the real-world empirical application, which ended up being the pretrial setting that I focus on in that paper, did I really realize that there was a lot of stuff in the theory that I hadn't fully though through that was important empirically.

Is there any particular result in your paper that you appreciate the most? Is there anything you are not yet satisfied with and may work further on in the future?

The result I got most excited about is that in this pretrial application, you can test whether choices are consistent with expected utility with any accurate belief under some weak assumptions about the decision maker's utility and the decision maker's private information. On the empirical side, when I actually applied that identification result to the data and found that a large fraction of judges are actually making decisions that are inconsistent with the model, I felt that was pretty exciting.

In terms of next steps, I really only provide some limited evidence of what we can say about the misspecified beliefs in the current paper – I provide some results that you can bound the extent of beliefs, and define accordingly overreaction and underreaction. But that could be consistent with many behavioral explanations of what's driving the prediction mistakes. Is it because judges fail to pay attention to all the evidence available to them? Is it because they overweigh or underweigh some piece of the information? It would be exciting to know whether we can use this sort of data pinning down whether choices are consistent with some type of behavioral mistakes or not. I think understanding the ways in which beliefs are biased will be helpful in understanding how decision makers like judges respond to the introduction of data-driven tools. If we can say something about why their beliefs are misspecified, perhaps it would suggest sensible policy implications.

What do you view as the biggest challenge as a graduate student? Was there anything you found difficult as a graduate student?

I think it is not particularly unique to me, but managing the transition from being a full-time student in the first two years to working on research was definitely challenging. I was very lucky to have great advisors who can help me out along the way, but that was certainly a challenge.

When learning about an interesting topic, it's easy to gather many relevant papers. Reading is rewarding but time is limited. How do you cope with the problem of having too much to read?

One piece of advice I got early on in graduate school is that you need to know when you have learned enough about existing literature to start working in it, but not enough to overly shape your thinking with how the literature approaches the problem. To answer the question, you may need to do something different. What I found is that, sometimes it is better to start earlier, and start to think about what you would do in this setting, and once you have that written out, and then going back to read more, rather than trying to read them all in one shot.

In retrospect, is there anything in your Ph.D. student life that you feel glad about and think you are doing particularly right? Is there anything you regret and wish you had been better informed about earlier?

One thing I am really glad I did is to start working on research relatively early on. When I started my third year, I had collaborations with faculty members and other graduate students. The way I learned about how to write papers is by working with faculty, with people who had published papers. That was very helpful. Something I still struggle with is when I remember being frustrated with myself during my Ph.D., not realizing that there are limited chunks of truly productive windows in a day or in a week. It's good to work hard, but you cannot slam your head on your desk to no end; that's not going to help you do better research.

Interview with Aislinn Bohren

In one of the tutorial sessions, Aislinn Bohren talked about two of her recent papers on the dynamics of discrimination and systemic discrimination. Dr. Bohren is an Associate Professor of Economics at the University of Pennsylvania, and she works actively on topics in microeconomics including discrimination, misspecified learning and information aggregation. Her work has both theoretical and empirical components.

Dr. Bohren's first paper, "The Dynamics of Discrimination: Theory and Evidence", introduces a dynamic dimension in the discussion of discrimination. In this paper, the empirical finding of dynamic reversal – the initial discrimination against one particular group ends up working in favor of this group – provides evidence for belief-based discrimination with incorrect belief. More importantly, this paper points out that dynamic reversal does not offset initial discrimination, and thus this sort of discrimination causes systematic under-rating for the initially discriminated group. The second paper, "Systemic Discrimination: Theory and Measurement", goes beyond the classic economic position of "direct discrimination" — holding all the other observables fixed to isolate the direct effect of group identity — and proposes the concept of "systemic discrimination", which demonstrates the pronounced indirect effects of earlier or contemporary discrimination.

In an interview after the winter meeting, Dr. Bohren gave further insight on both papers, and also shared some thoughts on doing economic research.

How did you start to study algorithmic fairness and issues of fairness in automated algorithms?

My dissertation research started with the project on the dynamics of discrimination. A lot of discrimination research in economics had focused on a very static question: in this period in time, is there discrimination, can we causally identify it, what are the sources, is it caused by some taste preferences or is it caused by beliefs. In a joint project with Alex Imas, we found a really neat online forum where you could test how discrimination evolves across time. Our intuition was discrimination may evolve especially if it's caused by beliefs. If people are discriminating at entry-level positions, then that could also affect discrimination at the promotion stage of the subsequent hiring stage.

The platform provides a publicly available reputation score, which is a summary of past performance on the site. So we used that to causally test how discrimination varies with positive past performance reviews. In that paper we found a discrimination reversal by gender. Looking at how math questions were evaluated, we generated the posts ourselves and then randomly assigned the quality of posts to each gender. We found discrimination against users with female names at the entry level, but users with female names are actually favored at the higher education level. Paired with the theoretical analysis, we showed that this is consistent with discrimination that stems from inaccurate beliefs.

In the course of that paper, we thought over the standard definition of discrimination used in economics, which identifies discrimination by holding fixed what's seen at that decision point and comparing populations with similar observables. But actually if a man and a woman who generate the same initial quality post receive different reputation scores after those posts, we should compare people who have similar quality posts, not people with similar reputation, because their reputation has already embedded some discrimination. That's what motivated our recent systemic paper, which was trying to broaden the definition of discrimination in economics to capture the systemic factors that seed the differences in the reputation aggregated at the decision point.

I think that's actually quite closely related to a lot of the algorithmic work in computer science. Within an algorithm, what direct discrimination would correspond to is differential treatments for users who have the same observables except for their group identity. That would be an algorithm that's explicitly using group identity to discriminate but the missing fact is that the algorithm might be trained on data that already has some discrimination baked into it. Even if you're using a group blind decision rule, say, a male user and a female user with the same observed variables are treated the same in the algorithm, still a male and a female with the same underlying productivity may be treated differently, because they have different observables that stem from discrimination in the past. So it's very closely connected to this idea that algorithms may lead to differential treatments across groups even if they do not explicitly discriminate against any group.

What would you say is a desirable goal for those conducting economic research on discrimination?

The first best would be to eliminate discrimination, but given that people often can't, any sort of policies to reduce discrimination, any decision node or any sort of way economists can provide evidence for a hypothesis that can reduce discrimination, will at least be moving things in the right direction. A key component of my paper with Alex Imas as well as Peter Hull is what we call total discrimination — the sum of direct and systemic discrimination. Essentially it needs to be defined against a reference point — that's a choice variable for the researcher. At the one extreme, your reference point is a constant and you can think of that is measuring all group-based disparities starting from birth or even before birth. On the other extreme, you could set that reference point as all the observables now and that would collapse things to the definition of direct discrimination. But there's a whole continuum of reference points in between. By choosing a reference point in between, it's not saying that the discrimination that occurred before doesn't matter or didn't occur; it's measuring discrimination that's occurred since this point.

Little bits of discrimination are added at each point in the pipeline and so trying to figure out where you can make decisions, then isolating the discrimination that's occurring along your decision nodes, can help researchers figure out effective policies to target that particular discrimination. Different policies will help eliminate discrimination in elementary school versus in higher university, so breaking it into different pieces will let you more effectively target a bit of it at a time. I really think this research can be useful, and people may not be aware of or have a way of thinking about systemic discrimination. So one of our goals of formalizing

it conceptually is to make people more aware that there may be bias baked into things you're looking at. Once you have a framework thinking about the systemic discrimination, it can help inform the public about a sense in which ways where discrimination may be baked into decisions.

Apart from economics, other social sciences are studying discrimination seriously as well, such as anthropology, sociology and psychology. How do researchers in these disciplines approach discrimination compared to economists?

Other fields like sociology and psychology have been thinking about this idea of systemic discrimination or algorithm fairness things from a broader perspective. So one of our goals of our systemic paper is to provide a framework to formally define and figure out how to measure and identify, using methods in economics, these more broad definitions of discrimination that other fields and other literature have been considering for a while, to try and bring economics up to speed in terms of thinking beyond just direct discrimination.

I also noticed that you have been doing research on misspecified learning. Do you have any suggestions for someone who wants to learn more about this topic, or any specific techniques or open questions that people should be familiar with?

Yeah, I think it's a relatively new literature, and I do think there's a lot of open questions. I'd say if you're interested in further reading, the literature review section in my paper has a lot of citations about other recent work, so I would start there and read through some of the other papers that have been written in the past. I think there's a lot of room for interesting applications like taking misspecified models and figuring out conceptually when they're relevant for particular markets and figuring out what sort of different predictions we can get or what empirical facts can be justified by the correct model in terms of belief updating with errors.

Is there any overarching theme or question that you constantly come back to in your research?

Broadly I'm really interested in how people learn from information. The assumption that people correctly interpret information and have rational expectations using Bayesian rules can be too strong. But once you relax that assumption, you, as a researcher, have a lot of degrees of freedom. It's really important to try and relax those models grounded on empirical work, observations of what's actually happening, but in a way that doesn't give you too much freedom. So I'm interested in providing theoretical foundations for these types of questions, and I also think a really important application is discrimination. In that line I'm also interested in taking these insights to more applied settings, like a discrimination setting, to see how they impact markets. That's one motivation for systemic discrimination; but, more broadly, I think that the idea of systemic discrimination is opening up a whole new door for really interesting research questions and so I'm hoping to also keep working in that area.

How would you distinguish good projects which truly contribute to the literature from those that are minor extensions from the existing literature?

It's hard to describe generally; it's more on a case-by-case basis. Basically, if you'd want to make a change to an existing model, you want to have good motivation for it. So you need to start with either some good psychological motivation or evidence for why this extension is relevant, rather than just saying like "Oh, you know, let me try and extend those elements in a random direction". You also need to generate new predictions that are testable and plausible from your model if you want to get a lot of mileage out of the new restriction. You want to make some sharp predictions and see it tie back to things that you actually see in markets.

What counts as a good paper for you? Of course, good papers should be written well and might use neat techniques. But other than that, is there any particular factor that you care about when judging a paper?

I like papers that are conceptually creative, papers that start with some evidence from the real world and then sort of conceptually push our boundaries and how we're thinking about things as economists. I think definitely there's technical papers that make important contributions too; but in terms of what I enjoy reading, I enjoy the conceptual creativity.

Could you share with us how you handle unproductive periods as a researcher?

I think if you're stuck on something, it's always good to work on a couple of projects. You don't want too many, because then you'll be too scattered; but also if you have only one and get stuck on it, you might just keep staring at the same thing, whereas if you have something else, you can just take a break and work on a different project for a week. Then you go back with a fresh perspective and can see the bigger picture. So have a couple of things you're working on, so that you don't be afraid to put something on for a few days. When you're just trying to do the same thing over and over, it's not working. For me, if I'm trying to think through something, I'll take a walk to figure a way to get around that issue.

REFERENCES

Bohren, J. A. and Hauser, D. N. 2021. Learning with heterogeneous misspecified models: Characterization and robustness. *Econometrica 89,* 6, 3025–3077.

Bohren, J. A., Hull, P., and Imas, A. 2022. Systemic discrimination: Theory and measurement. Tech. rep., National Bureau of Economic Research.

Bohren, J. A., Imas, A., and Rosenberg, M. 2019. The dynamics of discrimination: Theory and evidence. *American Economic Review 109,* 10, 3395–3436.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

Flanigan, B., Gölz, P., Gupta, A., Hennig, B., and Procaccia, A. D. 2021. Fair algorithms for selecting citizens' assemblies. *Nature 596,* 7873, 548–552.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*. PMLR, 1939–1948.

Heidari, H. and Kleinberg, J. 2021. Allocating opportunities in a dynamic model of intergenerational mobility. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 15–25.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.

Kearns, M. and Roth, A. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press.

Kleinberg, J., Mullainathan, S., and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Liang, A., Lu, J., and Mu, X. 2021. Algorithmic design: Fairness versus accuracy. *arXiv preprint arXiv:2112.09975*.

Rambachan, A. 2021. Identifying prediction mistakes in observational data. Ph.D. thesis, Harvard University.

Rambachan, A., Kleinberg, J., Ludwig, J., and Mullainathan, S. 2020. An economic perspective on algorithmic fairness. In *AEA Papers and Proceedings*. Vol. 110. 91–95.