Interview with Juba Ziani

YAHAV BECHAVOD
The Hebrew University of Jerusalem

Yahav Bechavod *is a fourth and final year PhD student in Computer Science at the Hebrew University of Jerusalem, where he is advised by Amit Daniely. In the winter of 2023, he will join The University of Pennsylvania as a postdoctoral researcher in the Computer and Information Science Department, where he will be hosted by Aaron Roth. His research interests are primarily in algorithms, machine learning, and game theory, and specifically in the areas of fairness in machine learning, online learning, and learning in the presence of strategic behavior, on which he has recently co-organized the "Learning and Decision-Making with Strategic Feedback" workshop at NeurIPS'21. He is honored to have been awarded The Israeli Council for Higher Education Postdoctoral Fellowship, The Apple PhD Fellowship in AI/ML, and The Charles Clore Foundation PhD Fellowship.*

In his winter meeting tutorial, Juba Ziani presented an overview of recent advancements in the rapidly evolving field of algorithmic fairness. Dr. Ziani is an assistant professor at the School of Industrial and Systems Engineering at Georgia Institute of Technology. His research focuses on the design of markets for data, data privacy, fairness in machine learning and decision-making, and strategic considerations in machine learning. Dr. Ziani's recent work explores the effects that opacity of algorithmic decision-making policies may have on the ability of individuals from different sub-populations to improve. In particular, it introduces a framework which is based on individuals learning about the policy from their peer networks, and demonstrates how information gaps between different sub-populations may result in an increased level of disparity. In an interview after the winter meeting, Dr. Ziani shared some of his thoughts on research on algorithmic fairness.

The field of algorithmic fairness has been expanding rapidly over the last several years. What are some central works that you would recommend to a researcher who is new to the field?

A first work I would recommend is *On the (Im)possibility of Fairness* [Friedler et al. 2016]. This paper basically tells you why you should do something about fairness, and why the naïve policy of being blind to race or gender or another sensitive attribute is not going to work. It shows how the same features might actually mean different things in the context of different backgrounds or sub-populations. A typical example that people use is how SAT scores have a different meaning depending on if you come from a wealthy background and can take the test multiple times and take an SAT preparation course or if you cannot afford these things.

Another paper I would recommend is *Fairness Through Awareness* [Dwork et al. 2012]. This paper introduces and formalizes an intuitive definition of fairness that many people have, which is that similar people should be treated similarly.

In addition, a lot of the space of fairness is trying to understand fairness at the

level of groups, and make sure that we treat different sub-populations fairly. There are many great works in this space, but to recommend only one, I would look at *Inherent Trade-Offs in the Fair Determination of Risk Scores* [Kleinberg et al. 2017]. It shows how guaranteeing fairness can be hard, by showing that except for very contrived settings, the three main statistical definitions of fairness people have been using are not compatible with each other. So, most of the time, you would have to only choose one. I think it is important that people understand this, because if you do not, and you choose to apply one definition of fairness without thinking about whether this is the right definition to apply, you might end up doing the wrong thing, and miss out on the "right" definition to use in this context.

Finally, the book *Fairness and Machine Learning: Limitations and Opportunities* [Barocas et al. 2019] is a good starting point to understanding how people have been thinking about, modeling, and addressing these issues.

One theme that was emphasized in the tutorial was algorithmic fairness in policy domains where individuals behave strategically. In this context, you mentioned a recent work of yours, *Information Discrepancy in Strategic Learning* [Bechavod et al. 2022], investigating the effects of information gaps between sub-populations in settings where individuals wish to act strategically, but lack the information to do so. Could you elaborate on some of the high-level takeaways from this work?

I think the first takeaway is that in many settings where you care about fairness, you do so because you are making high-stakes decisions about people. So you cannot simply imagine that people will not do anything if there is a chance they can steer the decision in their favor. People are naturally going to try to get to the positive side of these decisions. They are going to modify and work on their features to increase the chance of getting the loan that they need.

Another point is that a lot of the algorithms that are used in practice, for example for loan approvals or credit scores, are black-box proprietary algorithms. People may have some understanding or information about these models, but they don't know exactly how to tailor their responses to pass the classifier. Different people may also have different information about the classifier. One source of information is the decisions and predictions made on other people in the past. You can imagine getting this information from people similar to you, that are in your social or socioeconomic group. There could be great disparities as a function of that - if you are surrounded by wealthy people that get the loans they need, you probably have a very good sense of how this is going to work, but if your surroundings are people that typically haven't had access to those loans, then what are you going to do?

In the paper we look at the interaction between information asymmetries and strategic behavior, and aim to understand how they could lead to very different outcomes for people. One of the main messages of the paper is that even if the decision-making principal is benevolent and trying to push people to improve as much as possible, you can still have disparities across different groups and how much they are going to improve, depending on the information that they have. We even have situations in the paper where, by trying to optimize the improvement for the entire population, you are actually hurting the sub-population that doesn't have much information. The individuals in this sub-population actually change their

features in a way that doesn't improve their "true" label (think, for example, in a loan setting, their ability to repay the loan in a timely manner). One final takeaway here is that social welfare and fairness can be very different things, which I think a lot of people have realized by this point. Optimizing for social welfare doesn't necessarily mean you are doing something that is good for every sub-population.

Another theme in the tutorial was the effects of feedback loops - where new information is available only for applicants that have actually been accepted (e.g. approved for a loan, admitted to college). A recent work of yours, *Wealth Dynamics Over Generations: Analysis and Interventions* [Acharya et al. 2022] explores the effects of such feedback loops on the distribution of wealth across generations. Could you share some of the insights from this work?

The idea here is that wealthier populations generally have access to better opportunities in life. So, it is easier for them to keep wealth or gain wealth, and such effects can propagate over time. In order to study this, we consider a simple game-theoretic model that focuses on one salient element of such a feedback loop. The setting that motivates our paper is university admissions. What happens very often in this context, is that schools condition acceptance on doing well on tests such as the SAT or GRE. However, we know that these metrics can be biased; real data and empirical studies indicate that there are significant disparities in how well students from different socioeconomic and racial groups perform on standardized tests. In particular, wealth affects your ability to prepare for these tests. For example, when you take the SAT, part of doing well on the test is actually having the technical skills to answer the questions, but part of it is that you have been prepared to take this kind of test, that you know how to strategize, how to skip questions, how to take the test more efficiently; generally, this preparation is correlated with socioeconomic status. Further, there is an intrinsic disadvantage in that some people cannot afford to take the SAT more than once, while others take it several times, and only need to report the best score. So you could imagine two people who technically are similarly qualified for a university could possibly fare differently on the SAT. This effect is further amplified by policies such as legacy admissions, which explicitly give an advantage to individuals from wealthier populations that have historically had access to high quality college education.

In the paper, we show situations where the wealth update rule - how your wealth today is mapped to the wealth of the next generation - can have different fixed points in which people who start with low wealth are going to stay with low wealth, and people who start with high wealth are going to stay with high wealth. This is the first contribution of the paper. And then we study interventions. There are interventions of the form "what if we redesign the test?" "What if we increase or decrease the admissions threshold?" We also have interventions that are trying to increase the wealth of populations, by giving them subsidies, fellowships, funding that will make it easier for them to access educational opportunities. We study the effects of such subsidies and how one should give subsidies to people and when, in order to get rid of the feedback loops, and make sure that everyone converges to the same outcome in the long term.

Where do you see the field of algorithmic fairness going next? What are some questions that you think are the most interesting to explore?

To me, the higher level concern for the field to think about is to understand that algorithmic fairness is not just about algorithms; it's not just about adding a fairness constraint to my current algorithm to make the predictions more fair, because that may have unintended negative consequences in practice if you don't take the context into account. The question really is: what are the implications of your algorithm? How do they affect society? How do they impact people? Predictions are going to be translated into decisions, and we need to start moving away from trying to make predictions fair, and rather understand how these predictions are going to be used downstream and how they are going to affect people.

Another thing, which I touched on in the tutorial, is that decisions are not made in isolation. We face complex systems of decisions, and they interplay with and affect each other. I think at this point there is a pretty good understanding that even decisions that are fair in isolation are not necessarily going to be fair when you compose them with each other. I do not think, however, that it is well understood how to take that into account and how to deal with it outside of a few relevant papers, which means there is space for further understanding of these issues.

Going back to the feedback loops we talked about earlier, when we make high-stakes decisions about people and populations, there is a long-term impact. We are literally changing how these populations are going to be accessing opportunities in the future, and we have to take this long-term impact into account. I would argue that the goal of fairness is ultimately the long term, not the short term. We don't just apply affirmative action to account for past disparities, we also do so because by giving an opportunity to people we haven't given an opportunity so far, we are expanding access to educational opportunities to new groups that have been historically marginalized. The hope is that this will also expand access to opportunities and reduce disparities for future generations. When we think about our algorithms and the decisions we make, we really need to understand how those are going to shape the populations we'll face in the future.

We have to understand how the design of decision-making algorithms affects the data we collect about people. When you change your model, you change the way people are going to act. An example can be "we have never given a loan to people like you", so you think you are not going to get the loan so you won't even try. And then we are not going to collect data about these people. When we deploy an algorithm, we change the landscape in which the algorithm has been deployed, and we need to take it into account when we are thinking about fairness. These questions go back to incentives, game theory, behavioral modeling and economics, so I think that our community is exactly the right one to think about these types of questions.

In your earlier research, you were mostly focused on more "traditional" game-theoretic questions, while recently you have worked on fairness in machine learning. Can you share a bit about what made you interested in algorithmic fairness research?

I don't really think my research has changed so much over the years, and that's

for two reasons: first, the problems I care about in the space of fairness involve incentives and economic components, and use a lot of game-theoretic tools, so I'm still pretty much a game theory person. Second, even my earlier work, in more "traditional" game theory, was about ethical considerations and social impact. I started working on data markets, where we actually care about properly compensating people for their data instead of selling their data, and where you care about the privacy of the people whose data you are using in your computations. So in that sense I was always interested in those ethical concerns and issues.

I also care about it as someone who is from a minority background. I'm a minority in Algeria by being a Berber, and I'm a minority in France, the country I was born and grew up in, by being of Algerian descent. So I always cared about these issues that are related to fairness, diversity and inclusion, and it was natural for me to incorporate them in my research.

You are currently starting your second year as faculty at Georgia Tech Industrial and Systems Engineering. What is some advice that you can share from your first year as faculty, which could be helpful for those starting their first academic position?

Piece of advice number one: say no. Say no when there is too much to do in terms of reviews, projects, surveys and everything. And I would say it goes beyond saying no to other people, it's also saying no to yourself. At some point over this summer I was on five different projects, and I had these five other ideas, and I should really have realized that "now is not the right time to start working on these new ideas, stop reaching out to more people". I did a really poor job at this so I have too many things going on.

Another thing is that it's easy to get overwhelmed when you start as faculty. You get dropped into a new environment with many new considerations to take into account. It's an entirely new and different job. Before, the number one thing you had to care about was research; but when you become a professor, you have to do research, you have to teach, to do service, to apply for grants; there are so many things you have to do at the same time. My advice here is to take baby steps, one step at a time. It's okay if for one semester you focus mainly on only one of these things; for example I spent my first semester developing a new class on the algorithmic foundations of ethical machine learning, and frankly did not do much else. So my main advice is do one thing at a time, because otherwise you are just going to be overwhelmed.

REFERENCES

Acharya, K., Arunachaleswaran, E. R., Kannan, S., Roth, A., and Ziani, J. 2022. Wealth dynamics over generations: Analysis and interventions. *CoRR abs/2209.07375*.

Barocas, S., Hardt, M., and Narayanan, A. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. `http://www.fairmlbook.org`.

Bechavod, Y., Podimata, C., Wu, Z. S., and Ziani, J. 2022. Information discrepancy in strategic learning. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds. Proceedings of Machine Learning Research, vol. 162. PMLR, 1691–1715.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, S. Goldwasser, Ed. ACM, 214–226.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. 2016. On the (im)possibility of fairness. *CoRR abs/1609.07236*.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. 2017. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, C. H. Papadimitriou, Ed. LIPIcs, vol. 67. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 43:1–43:23.