Research Statement

Alice Huang

My research projects fall under two overarching themes. First, I use computational models to study how groups form beliefs. I am particularly interested in epistemic phenomena pertinent to our contemporary social discourse, such as misinformation, polarization and diversity. The first two papers in my dissertation on diversity and expertise belong to this category. Second, I am interested in how mathematical and empirical results about fairness constraints in machine learning should be interpreted, and how they relate to not only machine predictions, but also human decision making. The third paper in my dissertation on the fairness-accuracy tradeoff falls in this category.

Below I describe future research that I plan on undertaking, including projects that follow up on my previous work and new questions that I want to explore using similar methods.

From algorithmic fairness to testimonial justice

Putting our philosophical theory in dialogue with the formal literature on fairness, I hope to develop a more precise and rigorous account of testimonial justice tailored to the context of human credibility assessments.

One of the most well-known theorems in the literature on fair machine learning states that two of the most common statistical definitions of fairness cannot be jointly satisfied, except in trivial cases [Kleinberg et al., 2016]. The first criterion requires that the false positive and/or false negative rates be equal between two groups of different protected attributes (e.g., race, gender, marital status etc.). The second requires that the predictions match the actual outcomes equally well for both groups.

Stewart and Nielsen [2020] point out that this mathematical theorem applies to any classification problem, including human judgments of the credibility of others. Therefore, it is also impossible for us to simultaneously satisfy these two highly plausible criteria of fairness in our assessment of credibility. Given that failure to meet either criterion seems to be unjust, the theorem implies that testimonial justice (i.e., fair assessment of others' credibility) is mathematically impossible.

In light of this result, I want to explore whether we can locate the injustice of testimonial injustice elsewhere, such that norms of credibility assessment do not rely on criteria that cannot be jointly satisfy.

The Kleinberg et al. [2016] theorem concerns fairness constraints on predictions, but not the process by which an algorithm or a human arrives at the assessment. The focus on outcomes makes sense for machine learning algorithms, which are often unexplainable and designed to produce good predictions. But we might think that other criteria are more appropriate as formal definitions of fairness in the context of testimonial injustice. My project examines various other fairness criteria in the machine learning literature, notably those that focus on assessment-forming processes rather

than merely the results, such as fairness criteria aiming at removing protected attributes and their proxies from the assessment-forming process.

The silence of the moderate

Polarization is the phenomenon where two groups of agents form persistently divergent, and often radical, views, even when they are able to communicate with each other and receive new evidence. A lot of work has been done in epistemology to explain how mechanisms that are individually rational, such as Bayesian updating, can lead to polarization [Dorst, forthcoming, Nielsen and Stewart, 2021, O'Connor and Weatherall, 2017, Singer et al., 2019].

In my new project, I want to rethink social media strategies to mitigate polarization by exploring a related, but different question: what kinds of social incentive mechanisms can cause those with radical views to share their opinions more than those with moderate views?

In most formal models of epistemic communities, agents choose whether to solicit testimony from others, but always share their beliefs when solicited. This is not the case in reality. It is often a judgment call whether one should share what they think, or stay quiet. I plan to develop a model where agents decide to share their opinion when the expected cost of conflict is outweighed by its expected influence on the beliefs of others. I study conditions under which those with moderate opinions learn to withhold their views, whereas those with radical opinions learn to share more frequently. Under these conditions, our society will appear more polarized than it really is.

Studying these conditions is pertinent for several reasons. First, identifying conditions that can create an appearance of polarization might explain why social media seems to have amplified polarization, despite recent studies showing that the impact of social media on political polarization is grossly overrated [Guess et al., 2023a,b, Nyhan et al., 2023]. Second, the appearance of polarization might subsequently lead to real polarization, given the assumption made in previous work that we tend to trust similar others more. If those with moderate opinions refuse to communicate, this creates a gap in communication between those with radically opposing opinions. This project thus shed light on designs of social media strategies to mitigate polarization beyond simply exposing people to opposite opinions.

How to use track records

In my paper *Track Records: A Cautionary Tale* (currently R&R), I made a pessimistic conclusion about our ability to rely on track records to identify experts. The pessimistic view is based on the finding that, even in an idealized model where track records are fully available, the scientific community can fail to confer recognition on the most reliably scientists. It then follows that the laypeople cannot simply trust the opinions of the most prestigious scientists.

This pessimistic conclusion invites a further question: if fully transparent track records are not the solution, what it? As an extension of this paper, my new project explores how information about track records should be used.

Studying a different mechanism (reinforcement learning), Barrett et al. [2019] and Bruner and Holman [2022] have made more optimistic conclusions about using trade records to identify expert. It is, therefore, interesting to find the source of this disagreement. Comparing their models to mine, my hypothesis is that the problem uncovered in my model can be solved, if information about track records spreads slowly and locally, instead of being widely available immediately. In this new project, I make two modifications to the original model in my paper in order to verify or

falsify this hypothesis. The first modification involves making track records only *partially* available, and the second modification introduces errors in track records.

These modifications make the model more realistic. And whichever way the results turn out, these modifications will have interesting upshots. If they lead to improvements, we gain understanding on the kinds of contexts in which track records are helpful. If they do not lead to improvements, then the phenomenon uncovered in my previous paper is shown to be robust across a wider range of more realistic modelling assumptions.

References

J. A. Barrett, B. Skyrms, and A. Mohseni. Self-assembling networks. *British Journal for the Philosophy of Science*, 70(1):1–25, 2019. doi: 10.1093/bjps/axx039.

J. P. Bruner and B. Holman. Pooling with the best. In G. Ramsey and A. de Block, editors, *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*, chapter 2. University of Pittsburgh Press, Pittsburgh, PA, 2022.

K. Dorst. Rational polarization. *The Philosophical Review*, forthcoming.

A. M. Guess, N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, S. González-Bailón, E. Kennedy, Y. M. Kim, D. Lazer, D. Moehler, B. Nyhan, C. V. Rivera, J. Settle, D. R. Thomas, E. Thorson, R. Tromble, A. Wilkins, M. Wojcieszak, B. Xiong, C. K. de Jonge, A. Franco, W. Mason, N. J. Stroud, and J. A. Tucker. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, 2023a. doi: 10.1126/science.abp9364. URL https://www.science.org/doi/abs/10.1126/science.abp9364.

A. M. Guess, N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, S. González-Bailón, E. Kennedy, Y. M. Kim, D. Lazer, D. Moehler, B. Nyhan, C. V. Rivera, J. Settle, D. R. Thomas, E. Thorson, R. Tromble, A. Wilkins, M. Wojcieszak, B. Xiong, C. K. de Jonge, A. Franco, W. Mason, N. J. Stroud, and J. A. Tucker. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, 381(6656):404–408, 2023b. doi: 10.1126/science.add8424. URL https://www.science.org/doi/abs/10.1126/science.add8424.

J. M. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016. URL http://arxiv.org/abs/1609.05807.

M. Nielsen and R. T. Stewart. Persistent disagreement and polarization in a bayesian setting. *British Journal for the Philosophy of Science*, 72(1):51–78, 2021. doi: 10.1093/bjps/axy056.

B. Nyhan, J. Settle, E. Thorson, M. Wojcieszak, P. Barberá, A. Chen, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, S. González-Bailón, A. Guess, E. Kennedy, Y. Kim, D. Lazer, N. Malhotra, D. Moehler, and J. Tucker. Like-minded sources on facebook are prevalent but not polarizing. *Nature*, 620:1–8, 07 2023. doi: 10.1038/s41586-023-06297-w.

C. O'Connor and J. O. Weatherall. Scientific polarization. *European Journal for Philosophy of Science*, 8(3):855–875, 2017. doi: 10.1007/s13194-018-0213-9.

D. J. Singer, A. Bramson, P. Grim, B. Holman, J. Jung, K. Kovaka, A. Ranginani, and W. J. Berger. Rational social and political polarization. *Philosophical Studies*, 176(9):2243–2267, 2019. doi: 10.1007/s11098-018-1124-5.

R. T. Stewart and M. Nielsen. On the possibility of testimonial justice. *Australasian Journal of Philosophy*, 98(4):732–746, 2020. doi: 10.1080/00048402.2019.1706183.