# BRIDGING PREDICTION AND INTERVENTION PROBLEMS IN SOCIAL SYSTEMS

Lydia T. Liu (Princeton University),
Inioluwa Deborah Raji (University of California Berkeley),
Angela Zhou (University of Southern California),
Arvind Narayanan (Princeton University)

June 2nd – June 7th, 2024

## 1 Introduction

Over 60% of the US population lives in a jurisdiction that uses a RAT, or "Risk Assessment Tool", such as the COMPAS for pre-trial risk assessment [9]. "Millions of clinicians and more than 80% of clinicians in the United States alone" use MDCalc, the "most trusted medical reference for clinical decision tools." UT Austins GRADE admissions algorithm was developed 2012, and used for 7 gradute admissions cycles by the computer science department [6]. In short, data-driven machine learning (ML) and algorithmic tools are increasingly employed in critical decision-making processes – and the scale and impact of these system deployments are becoming increasingly meaningful in our every-day lives.

Machine learning methods, from the choice of objective function to feature selection to validation, are primarily designed to optimize out-of-sample predictive accuracy, and typically do not account for additional validity failures when predictions are used for intervention. As a result, many of the involved models are often framed as isolated *prediction problems* - with the goal of capturing relevant information about one sample of the population and extrapolating those learned patterns to make judgments of any another relevant sample within the same population. However, in reality, these ML models actually operate more like holistic policy *interventions* once deployed.

In practice, predictions are deeply informed and influenced by interactions between various stakeholders and existing infrastructure, and various deployment factors shape the impact of the model's use in existing decision-making processes, which in turn contributes directly to downstream consequences. For example, pre-trial risk predictions are used by correctional agencies for a range of decisions from placement to treatment planning to parole assessment [9], and student drop-out risk prediction tools in high schools influence the behavior of students, teachers and even the resources available to schools [15]. As such, predictions can have far-reaching consequences that changes the characteristics of populations. Therefore, myopically limiting the characterization of the model to *just* the quality of its predictions closes the door to the potentially richer discourse that could happen when we consider the model to be an intervention or part of an intervention pipeline.

Although some prior work [10] suggests that accurate prediction alone is sufficient to achieve policy objectives in certain interventional problems, other discussions emphasize the gap between predictions and decisions [22], and how the use of prediction in these cases is predicated upon critical causal assumptions [23]. Modern causal inference now integrates machine learning tools with considerations of causal validity

[3]. For instance, although there is a strong institutional history of policy program evaluation [2], which assesses the causal effects of interventions in the social setting, this lens has not been broadly applied to study the deployment of predictive algorithms as themselves interventions. Conversely, although recent technical work has studied algorithms that induce causal effects or a strategic response [8, 7], the methodological work thus far has not been empirically calibrated to real-world phenomena, and lacks adaptivity to the relevant policy contexts. Hence, combining these perspectives is crucial.

Even acknowledging the key role of causal inference in bridging predictions with interventions, there is a need to address existing critiques of causal inference and causal modelling in policy and social settings. For instance, empirical approaches in causal inference, such as randomized controlled trials, have known challenges with external validity and practical impact on policy interventions [24, 25]. On the theoretical side, recent work by [27] and [26] have pointed out significant conceptual flaws with interpreting social categories such as race and gender as causally manipulable variables (such as in a causal diagram [28]), arguing that that the causal modelling of social factors often lacks epistemic validity. A holistic validation pipeline of predictive interventions that hinges on causal validity must take these criticisms into account.

In this workshop, we re-visit the limitations of relying on the prediction paradigm description of machine learning to describe its design, development and influence within social systems. Offering statistical tools to analyze the impact of the model beyond its prediction outcomes, we will bring together an inter-disciplinary cohort to explore alternative views of adopting a more intervention-based lens to machine learning design, development and evaluation.

In summary, while recent work have highlighted the theoretical and practical importance of ML predictions in policy and social systems, there remains important open questions about the mechanisms connecting predictions to interventions, and consequently what new auditing and validity frameworks are necessary to ensure the beneficence of data-driven prediction and intervention in policy problems.

## 2 Workshop Overview

The workshop brought together thirty-five experts in machine learning, algorithmic accountability, causal inference, law, and quantitative and qualitative social sciences to present the latest developments and explore new directions in the field. There were fifteen talks, three thematic panel discussions, three working sessions, and two poster sessions highlighting student researchers during the workshop. The talks and sessions covered the following themes:

1. Human-Algorithm Considerations in Decision Making
2. Causal Inference and Decision-making
3. Limits of Individual Prediction
4. Implementation Science and Public Sector AI

**Cultivating connections across fields** One strength of the workshop was in convening interdisciplinary perspectives studying similar real-world systems. For example, Eli Ben-Michael discussed a randomized controlled trial of algorithmic recommendations, where Simone Zhang had actually conducted fieldwork and sociological observation of a different individual at the same site. Both talks discussed impacts of algorithmic tools from different points of view: statistical and sociological. However, both talks investigated substantively different questions: Simone focused on characterizing different effects/mechanisms of predictions, providing a sociological lens on computational perspectives of performative prediction co-developed by another participant, Juan C. Perdomo. Many perspectives also were informed by healthcare systems and health decision-making: Ashia Wilson's provocation (computer science) on the incompleteness of AUC alone for decision-making resonated with discussions from Daniel Malinsky (biostatistics) about interventional decision-making, while Mark Sendak (healthcare practice) provided a higher-level perspective on regulatory environments in health. Other domains – education, social welfare, etc – were also highlighted by participants as key themes throughout the workshop.

# 3 Presentation Highlights

## 3.1 Day 1: Human-Algorithm Considerations in Decision Making

The organizers (Lydia Liu, Deborah Raji, and Angela Zhou) opened the workshop with an overview of the research area and its motivations. The talk highlighted the ubiquity of automated decision systems, in particular, risk assessment tools in the public domain. The intellectual and practical challenges that motivated our particular agenda include the interdisciplinary nature of problems (involving fields such as computer science, statistics, sociology, law, and philosophy); institutional users (hence the need to consider institutional practices and default policy baselines); and the public-private deployment context (which requires accounting for a broader range of stakeholders, their incentives, constraints, motivations, and actions).

Day 1's talks shared the common theme of addressing human factors in algorithmic decision making, spanning several different application domains.

**Amanda Coston: Validity, problem formulation, and data-driven decision making** Amanda's talk examined frequently overlooked issues in problem formulation and validity that undermine the effectiveness of data-driven systems. Defining "validity" broadly as "does the model do what it purports to do?", she leveraged validity theory from the social sciences to create a taxonomy of challenges that jeopardize *construct*, *internal* and *external* validity in algorithmic decision-making. A guidebook that structures deliberation on early-stage problem formulation was co-designed in collaboration with public sector agency leaders and AI developers, frontline workers, and community advocates. The talk also addressed the prevalent issue of selectively missing data, specifically missing outcomes (when the decision downstream of prediction determines whether the outcome will be observed) and demonstrated a solution using causal inference techniques.

**Manish Raghavan: Reconciling human expertise and algorithmic predictions** Manish's talk addressed the observation that, in many contexts, algorithmic predictions performed comparably to human expert judgment. He highlights one out of many compelling reasons to keep humans involved in decision-making: humans can access information that algorithms cannot. The question is: do they make better decisions using that information? For instance, in medical settings, while algorithms assessed pathologies based on fixed data, doctors directly examined patients. Manish and collaborators built a framework to incorporate expert judgements to distinguish between instances that are algorithmically indistinguishable, with the goal of producing predictions that outperform both humans and algorithms in isolation. When they evaluated the methods in **clinical risk prediction contexts**, it was discovered that although algorithms outperformed humans on average, humans added valuable information in specific identifiable cases.

**Eli Ben-Michael: Does AI help humans make better decisions? A methodological framework for experimental evaluation** Eli's talk sought to answer a related yet entirely complementary question: do algorithmic predictions helps humans make better decisions *in deployment* as compared to a human alone or algorithmic predictions alone? His work with collaborators (including Kosuke Imai, another workshop speaker) introduced a new methodological framework that can be used to answer this question experimentally. They consider a single-blinded experimental design, in which the provision of AI-generated recommendations is randomized across cases with a human making final decisions. Under this experimental design, they showed how to compare the performance of three alternative decision-making systems–human-alone, human-with-AI, and AI-alone, and applied the method to the data from a randomized controlled trial of a pretrial risk assessment instrument [1]. In the **pretrial risk assessment context**, it was found that (1) AI recommendations do not improve the classification accuracy of a judge's decision to impose cash bail, (2) AI-alone decisions generally perform worse than human decisions with or without AI assistance, and (3) AI recommendations tend to impose cash bail on non-white arrestees more often than necessary when compared to white arrestees.

**Talia Gillis: Regulatory challenges for algorithmic fair lending** Talia's talk summarized debates about discrimination and AI regulation in the **context of consumer credit**, emphasizing the urgent need for robust regulatory frameworks. She discussed several legal regulatory challenges, including the gap between

traditional fair lending laws and current practices, such as the focus on input scrutiny in traditional discrimination law. Additionally, Talia highlighted the discrepancies between emerging regulatory frameworks and the realities of algorithmic decision-making, including her work with co regulating humans-in-the-loop and the generalizability of disparity metrics. Finally, Talia addressed current opportunities to operationalize fair lending compliance in ways that were previously limited.

**Student Poster Session 1** Student attendees, **Luke Guerdan, Angelina Wang, Ezinne Nwanko**, and **Michael Zanger-Tishler** presented lightning talks, followed by a poster session, where they discussed their work in detail and received feedback.

#### 3.1.1 Panel: "Clinical vs. Statistical Decision Making"

What is the value of statistical judgment? When can we not trust an algorithmic recommendation, and instead need to rely on trusting in the personalized experience and judgement of the decision-maker?

This panel focused on the tension between a "clinical approach" – allowing for a humanized, individual-centric decision-making – and an "actuarial" approach – engaging in more pattern-based statistical thinking and extrapolation – for decision-making [31].

We wanted the panelists to consider that tension under differing contextual and circumstantial lenses and, more importantly, discuss how to incorporate reflections of this tension into our considerations for what constitutes an appropriate algorithmic deployment.

**Ben Recht: Provocation** Ben's provocation took a historical view of psychoanalysist Meehl's pioneering contribution to this dilemma. He introduced the framing of the tension between an idiographic view (where each case is treated as unique) and a nomothetic view (where a population of cases is leveraged as collective evidence applicable to a particular case). Most interestingly, Ben concluded with Meehl's conclusion – that despite intuition on the importance of individual consideration, it seems that clinical prediction in many well-scoped pragmatic cases is typically overall less performant than statistical prediction, though this conclusion is clearly circumstantial. Ben concludes with an observation of the the tautology at play here – namely that, by stating that "bureaucratic technology wins at bureaucracy" (i.e. statistical prediction is best on average), we frame the human as a hazard or hero based on what is measurable, when the core of their value and potential for individualized consideration is best observed on case studies, scenario-based assertions that are ultimately singular and unverifiable. He concludes then by stating that rather than asking *if* human decision making are worse than algorithmic decision-making, we should ask *when* this is the case and identify the features of problems for which it is more appropriate to take on one approach over another.

**Jessica Hullman: Provocation** Jessica's provocation focused on the role of modeling decision-makers with precision and effectiveness. She noted the importance of task specification when setting up lab-like experiments in the HCI context. She further discussed how the lack of specificity in task design and modeling is what has interfered with more reproducible measurements for assessing human performance, as the humans involved simply adopt divergent views on the assumed goals of the task when the problem is not well-defined. She spoke also of how factoring in the variability of human participants, the context-specific cues of a task, and the differences between real world utility and abstract model scoring rules accounts for much of the differences between algorithmic and human decision-makers in a deployment environment.

**Aisha Wilson: Provocation** Aisha's provocation was oriented around the particular use of clinical and statistical reasoning for medical risk assessments. She discusses how in deployed decision-making contexts, the current measurements we use to describe our statistical thinking do not reflect the priorities and concerns of the human decision-makers involved. As an example, she discusses the intuitive presentation of the AUC-ROC curve to clinicians. She discusses the history of the use of the AUC curve but also goes into detail on how it does not reflect the intuitive process for many clinicians she interacts with, highlighting alternative metrics economic value and precision@K that more explicitly encode decision-making priorities.

**Panel Discussion** Much of the panel discussion focused on debating the role of statistical thinking in downstream decision-making, with panelists providing context for when statistical approaches were more or less appropriate. For instance, Aisha and Ben proposed that perhaps statistical methods were best suited for guiding institutional decision-making under administrative incentives, and perhaps less suited for more individualized scenarios. However, some in the audience raised that many decision-makers are by default operating as rule-following agents within administrative contexts, and are assessed against statistical parameters of success. Jessica in particular emphasized the value of modeling such decision-making contexts explicitly to understand how the constraints and incentives of the setting shape the decision-maker's responsiveness to the presented predictions.

The panel ended with Ben questioning the role of statistics for different means – although he admittedly took the stance that statistical thinking would likely trump clinical judgement under any notion of aggregate or average performance, he admitted that this evaluative framing was itself a statistical paradigm, and judged under the lens of different metrics, it might be possible to identify some interchangeable dynamic where one type of decision-making is more suitable than the other.

## 3.2 Day 2: Causal Inference and Decision-making

The second day began with a series of talks digging deep into 1) how causal inference frameworks can enable optimal decision-making via individual treatment rules and 2) how causal inference frameworks can be used to address statistical challenges and improve prediction models, whether by enforcing constraint desiderata on predictive models, elucidating a framework for interpretability analysis, or by elucidating and motivating new types of conditional independence assumptions used in combining datasets, with applications to inputing race/ethnicity information.

**Daniel Malinsky: Risk prediction vs individualized treatment rules for cardiovascular care decisions: a health equity perspective** Daniel's talk focused on comparing the more common risk-prediction paradigm for directing treatment (give a beneficial treatment to those at highest risk of a bad outcome) to the paradigm of *individualized treatment rules*, which give treatment to those who most benefit (have the highest heterogeneous treatment effect). He also related these two paradigms to current debates in medicine about the use of racial information in risk scores, outlining a position that the use of racial information for decision-making is permissible when it improves outcomes for those groups. He illustrated the differences between the paradigms and potential improvements in health equity in a large pooled observational study for cardiovascular risk. Some questions remain as to how this general guidance plays out with finite data and potential model-misspecification.

**Razieh Nabi: Mitigating Unfair Biases in Statistical Learning with a Focus on Causal Constraints** Razieh's talk focused on developing a general theoretical framework for causal estimation in statistical learning under causal constraints, such as regression with a no-direct-effect constraint of some variable (such as race). The work seeks predictive models that satisfy *causal constraints*, introducing interventional analysis and desiderata into the predictive paradigm. Her talk combined tools from causal inference, semiparametric inference, and constrained optimization to develop novel estimators, including extensive simulations and analytical insight from the optimization solutions as to how causally fair predictors differ from unconstrained predictors.

**Joshua Loftus: Model-agnostic explanation tools and their limitations** Joshua's talk leveraged causal analysis to study and improve the interpretability of pure predictive models. Current tools for interpretability such as partial dependence plots do not model dependence between features. The talk proposes to leverage causal models for the dependence between features to develop more refined notions of interpretability, introducing *causal dependence plots* which generalize partial dependence plots. The talk provides a causal interpretation of PDPs as estimating the *natural direct effect* of covariates $X$ on the predictor $\hat{Y}$. These new causal dependence plots study instead study *total dependence* by intervening on the predictor, using a causal model to change other predictors, and then plotting the new predcitions.

**Kosuke Imai: Estimating Racial Disparities When Race is Not Observed** Fairness and equity is a crucial concern in societal systems, and yet many practical systems in society do not measure race or other sensitive information that is necessary to assess disparities. Kosuke's talk introduces new methodology, motivated by an application project of Daniel Ho (another participant), which improves upon standard predictive models of race given proxy information (BISG, Bayesian Improved Surname Geocoding), by leveraging different causal assumptions. The new causal assumption posits that surname is conditionally independent of the name given (unobserved) race, location, and other characteristies, which sometimes is true by design if decision-makers are blinded to race information (as in the later lending example). This method can be interpreted as leveraging surnames as an instrumental variable for race, leading to direct estimation methods and other improved methods. A validation study on the North Carolina voter file (includes self-reported race) indicates improvement, and another substantive application.

### 3.2.1 Panel: "Normative Analysis in Causal Inference"

**Lily Hu: Provocation** Lily's provocation focused on the indeterminacy of discrimination analysis on the grounds of *similarly situated individuals*, given wide discretion in determining what makes individuals similarly situated individuals. Much of discrimination analysis proceeds by contrasting received outcomes of *similarly situated* individuals. Lily's argument is that there are normative determinations implicit in who is similarly situated: similarly meritorious, deserving, etc. In the language of causal inference, this corresponds to pointing out the indeterminacy of determining *overlap*. Overlap regions are those where different interventions might be observed and where causal contrasts can be valid. Overlap is typically determined by assessing variation in treatment conditional on covariates $X$. The key question is what justifies variable selection in causal inference, and what choice of covariates it is valid to assess overlap upon: conclusions about validity can vary widely with different choices of covariates.

**Alexander Tolbert: Provocation** Alexander's provocation first provided background on different perspectives on the philosophy of race and causal inference, depending on whether race (or aspects thereof) is or isn't a cause on manipulable or non-manipulable grounds. Alexander then points to his own previous argumentation advocating agnosticism about the "essence" of race: on the premise that it's justifiable to withhold judgment about causality in the presence of unobserved confounders, causal hypotheses about effects of race have many unobserved confounders, so agnosticism is reasonable. Alex also points out how *processes* of racialization and stratification can induce *positivity violations*. He then introduces perspectives on the philosophy of language and how these bear out on philosophical debates about race. Alex concludes on a puzzle: viewing racialization as inducing positivity violations can undermine the framework of making claims about the effects of race.

**Solon Barocas: Provocation** Solon's provocation focused on exploring the sociological context of what he calls "equivocal measures". The given formulation is that equivocal measures really measure the product of two terms: such as underlying incidence times reporting rate. This indeterminacy about the source of changes in equivocal measures allows institutions using them to interpret changes in them as proof of performance improvement or success, despite dubious context suggesting otherwise. A motivating example comprises of interpretations of COVID statistics, despite potential evidence otherwise regarding testing patterns and so on. Another example is that of predictive policing: bracketing aside the broader questions of whether to deploy them, they introduce examples of equivocal measures due to the selective labels problem of not measuring crime when not precision-allocating police to a location. This raises the question of whether such systems ought to be measured instead on harder-to-measure counterfactual terms: should we measure the counterfactual impacts of such systems on reducing crime that *otherwise would not have occurred*, regardless of measurement? Bracketing aside technical questions (some of which other participants at the workshop study), these examples raises interesting sociological questions as to the use of performance measures.

**Panel** The provocation speakers were joined by talk speakers Daniel Malinsky and Joshua Loftus. Angela Zhou moderated the discussion. This panel was focused more on discussing conceptual challenges and opportunities in causal inference methodology when studying important societal questions.

The first panel question was:

> The workshop focuses on societal systems. How does your varied disciplinary background and training highlight specific considerations when studying societal systems, which otherwise might not be apparently from purely mathematical or technical perspectives?

Panelists shared how their different disciplinary backgrounds highlight some specific considerations in the technical/theoretical/mathematical/statistical analysis of societal systems. Alex discussed some perspectives from traditional methods in philosophy, like normative reasoning and philosophy of science, and highlighted a specific opportunity: disaggregating categories in empirical practice, and bringing in more of the historical context as he did earlier in his provocation. Lily highlights how in the quantitative analysis of discrimination there is no disentangling of empirical and normative exercises: often various choices in empirical analysis can reflect normative positions. Dan highlights how analytic philosophy's extensive definitions have the benefit of always defining things relative to a certain context, which would be helpful for analysis of societal systems. Josh points out how his statistical training in high-dimensional analysis naturally leads to the importance of variable selection, given the high-dimensional realities of empirical practice.

Next we opened the panel discussion for questions from the audience, discussing various topics such as what are opportunities/benefits of formal methods in this setting, implications of historical processes of racialization for discrimination analysis, and whether we require conceptual consensus on the definition of race in order to proceed with other empirical and social analyses.

### 3.3 Day 3: Limits of Individual Prediction

Day 3's talks questioned the assumptions behind data-driven risk (or future outcome) prediction for individuals, across four distinct application domains: clinical decision making, life trajectory, student success, and child welfare. Together they identify several of the most pertinent limitations of the prediction paradigm in social systems.

**Berk Ustun: When personalization harms performance** Berk's talk focused on the issue of including personalized and sensitive attributes (such as sex, age, and HIV status) in machine learning models, to allegedly improve the prediction accuracy for diverse subpopulations. His work examined an unintended consequence of this practice—including personal attributes in the model can actually lead to worsened accuracy for minority groups. He discussed how these effects violate our basic expectations from personalization in applications like clinical decision support, and describe how they arise due to standard practices in algorithm development. The talk concluded by outlining practical approaches to mitigate these issues, including implementing "personalization budgets" to assess worsenalization and developing "participatory prediction systems" to allow individuals to consent to personalized predictions.

**Matthew Salganik: The origins of unpredictability in life trajectory prediction tasks** Matt's talk explored why some life outcomes are difficult to predict. This investigation involved in-depth qualitative interviews with 40 families sampled from a multi-decade longitudinal study. The sampling and interviewing process built on the earlier efforts of hundreds of researchers to predict life outcomes for participants in this study [30]. The qualitative evidence uncovered in these interviews, combined with a mathematical decomposition of prediction error, led Matt and collaborators to develop a new conceptual framework of predictability. The research suggests that unpredictability should be expected in many life outcome prediction tasks, even when using complex algorithms and large datasets.

**Juan Carlos Perdomo: The Relative Value of Prediction in Algorithmic Decision Making** Juan's talk discussed the modeling of a policy-making context in which predictive outcomes were to be balanced with a characterization of interventions and costs inherent to the deployment context.

**Shion Guha: Deconstructing Risk in Predictive Risk Models** Shion's talk walked through a detailed enthnographic exploration his research group had explored on the deployment of various child welfare risk assessment tools in Wisconsin. His team mapped out a network of involved stakeholders and their complicated interactions within a policy pipeline, discussing how various design details of the tools used impacted the decision-making outcomes of the involved stakeholders.

**Student Poster Session 2** Student attendees, **Ben Laufer, Hammaad Adam, Sayash Kapoor** and **Roshni Sahoo** presented lightning talks, followed by a poster session, where they discussed their work in detail and received feedback.

## 3.4 Day 4: Implementation Science and Public Sector AI

**Daniel Ho: Evaluating AI Systems in Government** In this unrecorded talk, Daniel overviewed applications of algorithms and artificial intelligence in the federal government, overviewing recent collaborations as well as discussing evaluation frameworks for performance assessment of AI and algorithms in the government.

**Simone Zhang: Social Mechanisms of Performative Prediction** Simone's talk began with vignettes and takeaways from her experiences in the court conducting fieldwork on how judges discussed the recommendations of risk assessment instruments. More broadly, her talk focused on unpacking sociological mechanisms for the notion of "performative prediction", i.e. predictions that in turn affect the probability of outcomes and events that they were designed to predict. While there is a great deal of technical work studying such predictions (some done by workshop attendees), Simone's talk focused on outlining sociological different mechanisms as to why people might make different choices to realize certain outcomes, and also introduced some surveys highlighting individual responses to a stylized version of a "failure-to-appear" risk assessment tool.

**Bryan Wilder: Learning treatment effects while treating those in need** Bryan's talk focused on optimal experimental design that also managed in-sample participant welfare, incorporating Pareto trade-offs between the optimal variance of the final treatment effect estimator, and giving treatment to those in most in need (as measured via baseline outcomes). With calibrated simulations from realistic social services data, he finds significant gains in in-sample welfare can be obtained with relatively little trade-off in final estimation accuracy.

### 3.4.1 Panel: "Understanding and assessing models in deployment"

We began the panel with two provocations (unrecorded, at the speakers' request) by **Mark Sendak** and **Suresh Venkatasubramanian** at the intersection of AI research, institutional politics, and the public good. Mark's provocation investigated the potential conflicts of interest and equity issues that arise within multi-stakeholder health AI initiatives, and the current evolving landscape of public-private collaborations. Suresh's provocation explored the challenges of integrating algorithmic fairness research into policy documents for AI, highlighting the need for specificity, local context, and adaptability in policies to counteract the responses of commercial parties. He also emphasized the importance of focusing on developing processes rather than just algorithms.

The third and final provocation of this session was delivered by **Marissa Gerchick**, a data scientist at the American Civil Liberties Union (ACLU). Marissa presented evidence for ACLU's recent FTC complaint against major hiring technology vendor Aon for marketing their hiring tests as bias free. Research on racial disparities in Automatic Speech Recognition (ASR) systems [29] was cited to support the complaint. Marissa emphasized that the problems with the hiring software extended beyond the algorithms themselves. The talk also discussed the future of digital discrimination in hiring, advising employers to avoid tools that carry a high risk of discrimination.

**Panel Discussion** The three provocation speakers were joined in discussion by talk speaker and fourth panelist Daniel Ho. Lydia Liu moderated the discussion and posed the first question: *what are the different roles of research and researchers in addressing the complex problems surfaced in the provocations, and identifying under-researched areas?* Panelists highlighted the need for more 'public airing' when policies contradict research, emphasizing the role of civil society in applying pressure and providing alternative viewpoints. They also noted the undervalued role of academic research to broaden the vision of questions asked, understand the population impacted by algorithmic harm, and expand what is technically possible to address these issues.

While third-party auditing plays an important role in the AI accountability landscape, some speakers emphasized the need for auditing accountability, citing the pitfalls of an auditing industrial complex. Other participants advocated for transparency to enable broader participation beyond just entrenched auditors and auditing bureaucracies. Subsequent discussions touched on the importance of separating (internal) quality assurance from auditing and the need for standardized overseers within organizations for AI tools and products.

The panel also addressed regulation and legal frameworks, noting the challenges agencies like the FDA face in regulating algorithmic decision systems (ADS), such as determining thresholds for re-testing and the delegation of responsibilities to developers. There was a discussion on affirmative delegation for the least discriminatory alternative (LDA) and the need for anti-discrimination doctrine to keep pace with technology, often driven by litigation. The discussion concluded with the observation that anti-discrimination remains a key tool for identifying and dismantling invalid or flawed algorithmic systems.

# 4 Scientific Progress Made

An important objective of the workshop was to collaborate on a larger whitepaper including 1) survey the current state of the field and 2) prescriptive angles and paths forward for the field. We held three working sessions. In the first working session, we fostered discussions by breaking into subgroups covering: *model design, evaluation science*, and *implementation science*. In the second working sessions, we sought consensus on the recommendations raised in the initial working session, and further organized the groups' viewpoints pertaining to the prescriptive angles, which included *alternatives to prediction*, *model improvement and desiderata*, *holistic understanding and assessment of models in deployment*, and *sustainable governance and maintenance of ADS*. The third and final working session focused on the development of prescriptive angles. The workshop organizers and interested participants will continue to draft and refine the white paper after the workshop, and plan to publish the results by the end of 2024.

# References

[1] Imai, K., Jiang, Z., Greiner, D., Halen, R. & Shin, S. Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *Journal Of The Royal Statistical Society Series A: Statistics In Society*. **186**, 167-189 (2023)

[2] Glennerster, R. & Takavarasha, K. Running randomized evaluations. *Running Randomized Evaluations*. (2013)

[3] Athey, S. & Imbens, G. Machine learning methods for estimating heterogeneous causal effects. *Stat*. **1050**, 1-26 (2015)

[4] Ledford, H. Millions of black people affected by racial bias in health-care algorithms. *Nature*. **574**, 608-610 (2019)

[5] Green, B. & Chen, Y. The principles and limits of algorithm-in-the-loop decision making. *Proceedings Of The ACM On Human-Computer Interaction*. **3**, 1-24 (2019)

[6] Waters, A. & Miikkulainen, R. Grade: Machine learning support for graduate admissions. *Proceedings Of The 25th Conference On Innovative Applications Of Artificial Intelligence*. (2013)

[7] Kleinberg, J. & Raghavan, M. How do classifiers induce agents to invest effort strategically?. *ACM Transactions On Economics And Computation (TEAC)*. **8**, 1-23 (2020)

[8] Hardt, M., Megiddo, N., Papadimitriou, C. & Wootters, M. Strategic classification. *Proceedings Of The 2016 ACM Conference On Innovations In Theoretical Computer Science*. pp. 111-122 (2016)

[9] Brennan, T., Dieterich, W. & Ehret, B. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice And Behavior*. **36**, 21-40 (2009)

[10] Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. Prediction policy problems. *American Economic Review*. **105**, 491-95 (2015)

[11] Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*. **5**, 153-163 (2017)

[12] Shapiro, A. Reform predictive policing. *Nature*. **541**, 458-460 (2017)

[13] Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. **366**, 447-453 (2019), https://science.sciencemag.org/content/366/6464/447

[14] Coston, A., Mishler, A., Kennedy, E. & Chouldechova, A. Counterfactual risk assessments, evaluation, and fairness. *Proceedings Of The 2020 Conference On Fairness, Accountability, And Transparency*. pp. 582-593 (2020)

[15] Liu, L., Wang, S., Britton, T. & Abebe, R. Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings Of The National Academy Of Sciences*. **120**, e2204781120 (2023)

[16] Liu, L., Barocas, S., Kleinberg, J. & Levy, K. On the actionability of outcome prediction. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **38**, 22240-22249 (2024)

[17] Jordan, M. & Mitchell, T. Machine learning: Trends, perspectives, and prospects. *Science*. **349**, 255-260 (2015)

[18] Hofman, J., Watts, D., Athey, S., Garip, F., Griffiths, T., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M., Vazire, S. & Others Integrating explanation and prediction in computational social science. *Nature*. **595**, 181-188 (2021)

[19] Watts, D., Beck, E., Bienenstock, E., Bowers, J., Frank, A., Grubesic, A., Hofman, J., Rohrer, J. & Salganik, M. Explanation, prediction, and causality: Three sides of the same coin?. (OSF Preprints,2018)

[20] Narita, Y. & Yata, K. Algorithm is experiment: Machine learning, market design, and policy eligibility rules. *ArXiv Preprint ArXiv:2104.12909*. (2021)

[21] Haushofer, J., Niehaus, P., Paramo, C., Miguel, E. & Walker, M. Targeting impact versus deprivation. (National Bureau of Economic Research,2022)

[22] Athey, S. Beyond prediction: Using big data for policy problems. *Science*. **355**, 483-485 (2017)

[23] Lundberg, I., Brand, J. & Jeon, N. Researcher reasoning meets computational capacity: Machine learning for social science. (SocArXiv,2022,5), osf.io/preprints/socarxiv/s5zc8

[24] Stephenson, J. & Imrie, J. Why do we need randomised controlled trials to assess behavioural interventions?. *BMJ*. **316**, 611-613 (1998), https://www.bmj.com/content/316/7131/611

[25] Deaton, A. & Cartwright, N. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*. **210** pp. 2-21 (2018)

[26] Hu, L. & Kohler-Hausmann, I. What's sex got to do with machine learning?. *Proceedings Of The 2020 Conference On Fairness, Accountability, And Transparency*. pp. 513-513 (2020)

[27] Kohler-Hausmann, I. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*. **113** pp. 1163 (2018)

[28] Pearl, J. Causal diagrams for empirical research. *Biometrika*. **82**, 669-688 (1995)

[29] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J., Jurafsky, D. & Goel, S. Racial disparities in automated speech recognition. *Proceedings Of The National Academy Of Sciences*. **117**, 7684-7689 (2020)

[30] Salganik, M., Lundberg, I., Kindel, A., Ahearn, C., Al-Ghoneim, K., Almaatouq, A., Altschul, D., Brand, J., Carnegie, N., Compton, R. & Others Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings Of The National Academy Of Sciences*. **117**, 8398-8403 (2020)

[31] Dawes, R., Faust, D. & Meehl, P. Clinical versus actuarial judgment. *Science*. **243**, 1668-1674 (1989)