Research Statement

Debmalya Mandal
Data Science Institute
Columbia University
dm3557@columbia.edu

1 Overview of Research Interests

I work on problems in AI and statistical machine learning, with an emphasis on theory and modeling, and with applications focused on high-stakes, societal problems. I am particularly interested in how AI/ML systems can be integrated in society where where the decisions directly impact people's lives. The success of AI in diverse areas like machine translation to cancer diagnosis has led to its adoption in various, high-stakes societal problems. Examples include predictive policing, criminal justice [KLL+17], and also healthcare [Top19]. However, current AI systems cannot be used reliably across a variety of societal applications, and there remain several major challenges in the development of human-centered AI systems.

- **Elicitation**: Modern deep learning systems require millions of accurately labelled training instances [Mar18]. Acquiring such datasets is costly and often overlooked in the design of AI systems. We will need to make data collection an integral part of design and construct mechanisms to obtain high quality feedback from human labellers or annotators.
- **Aggregation**: As AI systems are increasingly being used for making high-stakes decisions in society, we need to ensure that such systems reflect appropriate values and ethics. In many situations, it is not a priori clear what are the right trade-offs [ADK+18], and one fruitful direction is to aggregate people's opinions about the various trade-offs AI systems face.
- **Evaluation**: Before an AI system is deployed in society, we need to evaluate whether its impact is positive or negative. Often such systems will interact with humans over long periods of time, and we need to develop methods to estimate the long-term effects of such systems.
- **Bias**: Finally, existing machine learning algorithms are often trained on biased datasets, and the resulting models exhibit discrimination against misrepresented groups [BG18]. Therefore, we will need to ensure that any ML algorithm is unbiased and robust before deploying such an algorithm in society.

In the following, I describe my research in several directions to address the four themes above – *peer prediction* to solve the problem of information elicitation, *voting* for the problem of information aggregation, *causal inference* for the problem of policy evaluation, and finally the design of *fair and robust* algorithms to reduce bias in AI systems.

2 Peer Prediction

Peer prediction studies the elicitation of information that cannot be verified, either because there is no objectively correct answer or because the answer is too costly to acquire. This problem arises in diverse contexts from grading peer assignments to reviewing services like a restaurant or a hotel in an online platform. As a concrete example, suppose a restaurant opens a page on Yelp to inform people about their various services. Such a page also contains feedback and ratings from users. It is crucial this information be accurate so that people are well informed. It can be costly for users to provide such feedback, and peer prediction mechanisms incentivize the users to invest effort and report truthfully by providing appropriate payments to users (in the form of money, points, or similar).

Existing peer prediction mechanisms ignore the fact that users providing feedback may be quite different in the way they think about the world. For example, users reviewing a particular business on a Facebook page may belong to various communities and have significantly different opinions and tastes. Our work [AMPS17] developed the first informed truthful peer prediction mechanisms for heterogeneous users. We achieved truthful reporting by first identifying the cluster of a user and then using appropriate scoring functions when two users from different clusters are paired together.

In recent years, the field of peer prediction has seen significant improvement in developing algorithms to handle various aspects of the problem like agent heterogeneity, task heterogeneity [MLP+16] etc. In a recent work [MGP20], we demonstrated the effectiveness of peer-prediction based methods in long-term forecasting of geo-polictical events. Through a large-scale experiment on Amazon Mechanical Turk we were able to show that providing feedback based on peer prediction mechanisms has a significant effect in increasing user engagement with the forecasting platforms. Moreover, hybrid schemes that combine traditional scoring rule based method with peer prediction methods, are the best in terms of accuracy of the forecasting platform.

An interesting research direction is to use peer prediction methods to train machine learning algorithms with noisy data. In fact, [LG20] used peer prediction methods to design new loss functions that improve learning with noisy labels. In summary, I think peer prediction methods should be adopted in both the design of AI and in its applications in various domains like online education, and rating systems.

3 Voting

Social choice theory studies the aggregation of individual preferences towards a collective decision. A canonical problem is the design of voting rules that aggregate preferences over a list of candidates and select the winning candidate. One drawback of standard voting methods is that it only asks users to provide a rank over the set of possible alternatives. In a platform with a large number of alternatives like a participatory budgeting platform, the performance of such voting rules can be poor, when measured in terms of utilitarian social welfare. However, if the users are asked to reveal additional cardinal information, we can significantly improve the quality of the selected alternative, measured in terms of its social welfare. In a sequence of recent papers [MPSW19, MSW20], we have characterized the trade-off between the achievable social welfare and the communication complexity of voting rules, which measures the amount of information the voters must convey to the voting rule. Our upper bound proposes new voting rules based on sketching algorithms, which we hope to be useful for settings with large number of alternatives. On the other hand, our lower bound makes interesting connections with the rich literature on communication complexity, which may be of independent theoretical interest.

As AI systems are increasingly being used in various societal contexts, we need to ensure that such systems reflect appropriate societal values. In many situations, the right choices are often unclear, and social choice theory provides a promising framework to incorporate and aggregate people's opinions. Following immense success of direct democracy platforms like participatory budgeting, I think more AI systems should employ voting as a means to understand trade-offs in various decisions.

4 Causal Inference

Consider the problem of determining the causal impact of a policy, e.g. what is the effect of a marketing campaign on a user's purchasing behavior, or how does offering discounts affect the number of trips taken by a user on a ride-sharing platform? There are two main challenges in this problem – (1) The pool of users visiting a large platform is inherently heterogeneous, and (2) such treatments are often applied sequentially for a large number of rounds. Historically, most datasets have been too small to uncover such heterogeneity in treatment effects. With the advent of large-scale experiments on online platforms, and improvement in computational and statistical methods, we have started to uncover the heterogeneity in treatment effects.

Although there have been significant attempts to utilize big data for discovering heterogeneity in treatment effects [Ath17], most of these attempts only considered cross-sectional data where the

treatment is applied only for a single time. In a recent work [MP19], we extend the *Marginal Structural Models* (MSM) [Rob00], the most widely used method to estimate the causal quantity of interest when the subjects receive treatments over multiple periods of time. We propose a new form of MSM which models the potential outcomes using a three-dimensional tensor, where the three dimensions correspond to the agents, time intervals, and set of possible histories. Unlike the traditional MSM, we allow the dimensions to increase with the number of agents and time intervals, which lets us account for the heterogeneity of the users and effects across a large number of time intervals. We show how to efficiently estimate the parameters of our model and show how different types of causal impacts can be derived from our estimator.

It is expected that significant parts of the economy, including finance, and consumer markets will eventually be automated with AI systems. Such increased automation could either lead to an overall increase of welfare of the entire society, or could increase the existing inequalities in our society. [BM14] raises the question of exploring various policies for incorporating AI systems in our society. In order to compare different policies, we need to design long-term experiments, and develop new tools to understand long-term impacts of AI in society.

5 Algorithmic Fairness

Over the past couple of years, there have been significant efforts in the machine learning community to develop fair algorithms. In most settings, the main goal of such a fair algorithm is to equalize some statistical metric of fairness over different groups of the population. However, the literature on fairness has largely ignored the design of fair and robust classifiers. In a recent work [MDJ+20], we found that the existing fair classifiers become unfair even if we slightly perturb the training distribution. This led us to study the design of fair classifiers that are robust to perturbations in the training distribution.

In our work, we develop classifiers that are fair not only with respect to the training distribution, but also for a class of distributions that are weighted perturbations of the training samples. We formulate a min-max objective function whose goal is to minimize a distributionally robust training loss, and at the same time, find a classifier that is fair with respect to a class of distributions. Experiments on standard machine learning fairness datasets suggest that, compared to the state-of-the-art fair classifiers, our classifier retains fairness guarantees and test accuracy for a large class of perturbations on the test set.

Algorithmic fairness is a growing field with many interesting open questions. I am particularly interested in designing algorithms that provide individual fairness guarantees. Instead of group fairness constraints discussed above, such algorithms provide stronger guarantees and ensure that similar individuals should be treated similarly [DHP+12]. However, such stronger guarantees of fairness come with more challenges – (1) It is generally impossible to provide individual level guarantees without a high loss in performance, and (2) The individuals and their perceptions about the decisions of the algorithm change with time. Therefore, we need to develop a better understanding of the trade-offs between fairness and performance, and ensure that the algorithms are robust before they can be reliably deployed in society.

References

[ADK+18] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The Moral Machine Experiment. *Nature*, 563(7729):59, 2018.

[AMPS17] Arpit Agarwal, Debmalya Mandal, David C. Parkes, and Nisarg Shah. Peer Prediction with Heterogeneous Users. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 81–98, 2017.

[Ath17] Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[BM14] Erik Brynjolfsson and Andrew McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. WW Norton & Company, 2014.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[KLL+17] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2017.

[LG20] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. *Proceedings of the International Conference on Machine Learning*, 2020.

[Mar18] Gary Marcus. Deep Learning: A Critical Appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[MDJ+20] Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette M Wing, and Daniel Hsu. Ensuring fairness beyond the training data. *Advances In Neural Information Processing Systems*, 2020.

[MGP20] Debmalya Mandal, Radanovic Goran, and David C Parkes. The effectiveness of peer prediction in long-term forecasting. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 2160–2167, 2020.

[MLP+16] Debmalya Mandal, Matthew Leifer, David C Parkes, Galen Pickard, and Victor Shnayder. Peer Prediction with Heterogeneous Tasks. *NIPS 2016 Workshop on Crowdsourcing and Machine Learning*, 2016.

[MP19] Debmalya Mandal and David Parkes. Weighted tensor completion for time-series causal inference. *arXiv preprint arXiv:1902.04646*, 2019.

[MPSW19] Debmalya Mandal, Ariel D Procaccia, Nisarg Shah, and David Woodruff. Efficient and thrifty voting by any means necessary. In *Advances in Neural Information Processing Systems*, pages 7180–7191, 2019.

[MSW20] Debmalya Mandal, Nisarg Shah, and David P Woodruff. Optimal communication-distortion tradeoff in voting. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 795–813, 2020.

[Rob00] James M Robins. Marginal Structural Models Versus Structural Nested Models as Tools for Causal Inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000.

[Top19] Eric Topol. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Hachette UK, 2019.