Research Statement

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

1 Overview of Research Interests

I work on problems in AI and multi-agent systems, with an emphasis on theory and modeling, and with applications focused on high-stakes, societal problems. I am particularly interested in how AI systems can be integrated in society where where the decisions directly impact people's lives. The success of AI in diverse areas has led to its adoption in many high-stakes societal problems, including 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.
- **Learning**: Once an AI system is deployed in society, it has to continually interact with end-users, and learn their preferences. This situation often appears in recommender systems [LMJ20], where the value of matching a user to a service is a priori unknown. Therefore, we need to ensure that such systems incorporate learning of user preferences.
- **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, *bandits and matching* for the problem of learning preferences, 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 in an online platform. As a concrete example, consider the problem of ensuring appropriate feedback on a restaurant's Yelp page. 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. Traditionally, social choice theorists study settings where individuals have independent, subjective preferences. However, several modern domains such as crowdsourcing and social networks require modeling dependent and uncertain preferences. In our work [MP16], we study voting with dependent preferences, and show that a large class of voting rules become incentive-aligned exponentially fast in the number of voters when the beliefs of the voters are *positively-correlated*. Such positively-correlated beliefs often show up in standard rank-order models, and our results suggest the benefits of voting in aggregating preferences under these models.

Another 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.

Another drawback of existing voting rules is that, when the number of alternatives is large, they require many votes to converge to the correct answer. In a recent work [HMSS21], we show how to alleviate this problem by asking voters' additional prediction questions about the opinions of other voters. Through a large-scale experiment on Amazon Mechanical Turk we were able to show that the combination of votes and prediction reports significantly outperforms the classical voting rules on questions from different domains.

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 Online Learning

Multi-armed bandits have been used extensively to model sequential decision making. However, they usually ignore that services (i.e. arms) are often unavailable for some number of rounds once they are allocated (i.e. pulled). This problem appears frequently in cloud-computing where a resource is often blocked after allocating it to a task. In a recent work [BCMTT20], we model this problem through *adversarial blocking bandits* where the rewards and the blocking lengths of the arms can vary arbitrarily over time. Since there is no fixed best arm in this context, our main contribution is establishing the right benchmark. When the rewards and the blocking lengths are known, we show that finding the optimal policy for this problem is NP-hard and we construct a greedy policy that is a constant factor approximation of the optimal policy. Then we design a learning algorithm that has sublinear regret when measured with respect to the greedy benchmark.

For most real-world settings, AI systems often interact with multiple agents, and the systems needs to learn the preferences of different users. One example of such a multi-agent setting is online matching where the value of matching a user to a service is apriori unknown [LMJ20], and the system must learn how much value a user receives when assigned to a particular service. One drawback of the recent literature on learning matching is that they don't consider the fact that certain services could get blocked once assigned to a user. In a recent work [BCMTT21] we consider a setting where allocating a service to a user blocks that service for a certain number of rounds. Since there is no unique matching that can be applied repeatedly over time, we first characterize the offline benchmark, the policy that should be applied even when all the rewards and blocking lengths are known. We then derive a multi-agent learning algorithm that has per-agent logarithmic regret with respect to the offline benchmark.

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.

6 Future Research Directions

Here I outline some immediate directions for future work and outline several broad research directions I would like to pursue over the next couple of years.

Individual Fairness

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, the existing notion of individual fairness requires a metric to compare different individuals and the metric might be different for different contexts. One way to determine such metric would be to elicit different experts' opinions and aggregate them in a meaningful way, and in this regard, I think both peer prediction and voting will have important role to play.

Individual fairness is often a contextual problem and different contexts require different metrics to evaluate fairness. It might be impossible in general to automatically determine such a metric, and one idea would be to solicit opinions from experts and aggregate them. However, even experts differ in their judgements and we have to resort to using social choice theory to come up with an appropriate aggregation rule. Current approaches only aggregate pairwise comparisons over two alternatives (fair or unfair) which can be solved by majority rule [JKN+19]. However, going beyond this paradigm requires appropriate formalization through the lens of social choice theory.

Individual fairness metric is often complicated and might be hard for an expert to state explicitly. Metric elicitation provides an alternate framework to elicit such measures through queries about classifier outcomes, and has been applied to elicit group fairness comparisons [HNK20] and individual fairness metrics [Ilv20]. However, our ultimate goal is to automate this elicitation process and deploy it on a larger scale. This requires often asking experts many queries and we want to incentivize them to invest effort and report accurate metric. The method of Peer prediction was developed precisely for this purpose and will definitely have an important role to play in automating the design of fair algorithms.

Robust Optimization for Fairness

A significant drawback of existing fair classifiers is that they guarantee fairness on a fixed training distribution, and such guarantees do not hold when the distribution is slightly perturbed. Our recent work [MDJ+20] shows how to construct group-fair classifiers that are robust to perturbations in the training distribution. Several recent papers have also considered the problem of designing fair classifiers under adversarial robustness. However, I believe, both distributional and adversarial robustness are not appropriate for algorithmic fairness. To give an example, consider the problem of determining whether an individual will recidivate or not. A plausible hypothesis is that $P(\text{race})$ or $P(\text{gender}|\text{race})$ will change from one county to another, but $P(y|\text{education-status}, \text{age})$ should remain the same.

This suggests that there are certain independent mechanisms (i.e. factors) that are invariant and the remaining factors might change from one domain to another. So, one should build a fair classifier that is robust to changes only in the latter kinds of factors. Such notions of independent mechanisms are formalized by [SLB+21], and I believe, should help mitigate the significant drop in accuracy due to ensuring robustness. Of course, the main problem is identifying the factors that don't change. In general, it might be impossible to automatically detect such factors because of samples from a limited number of domains and we should resort to human feedback. However, when the number of variables is too large, the number of possible factors could also be very large. So we need to carefully design queries to ask human experts, in order to approximately cover all the factors and determine their nature.

Multi-Agent Systems

As most real-world AI systems have multiple agents at stake, such systems must incorporate interacting with multiple agents over time. Despite many recent progress in multi-agent learning systems, there are several challenges to overcome. One of the major challenges include various constraints in the allocation algorithms e.g. blocking, safety, fairness etc. Our recent work [BCMTT21] on blocking bandits model captures how to model blocking in a matching market. However, this is a simple model and it would be interesting to extend this model to consider broader class of allocation problems and more general types of constraints.

Furthermore, our model assumes that the agents are static and their preferences do not change over time. In practice, most multi-agent systems are dynamic in the sense that agents' preferences evolve over time. Therefore, a more general model is to consider a setting where the preferences of the agents evolve according to a Markov Decision Process (MDP) and our goal is to learn a joint policy that maximizes some long-term objective. Although this setting falls under the framework of multi-agent reinforcement learning (MARL) [ZYB21], there are several challenges in applying the existing algorithms. First, standard MARL setting doesn't consider the problem of incentivizing the agents to explore and learn their private models. Second, constraints like blocking, safety, fairness etc often introduce non-standard benchmarks as the Bellman equation cannot be applied. Overall, I think the general framework of learning coordinating policies for multi-agent systems under various constraints have many interesting research problems and will be an exciting area for future research.

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.

[BCMTT20] Nicholas Bishop, Hau Chan, Debmalya Mandal, and Long Tran-Thanh. Adversarial blocking bandits. In *Advances In Neural Information Processing Systems*, 2020.

[BCMTT21] Nicholas Bishop, Hau Chan, Debmalya Mandal, and Long Tran-Thanh. Sequential blocked matching. *arXiv preprint arXiv:2108.00073*, 2021.

[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.

[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.

[HMSS21] Hadi Hosseini, Debmalya Mandal, Nisarg Shah, and Kevin Shi. Surprisingly popular voting recovers rankings, surprisingly! *The Thirtieth International Joint Conference on Artificial Intelligence (Forthcoming)*, 2021.

[HNK20] Gaurush Hiranandani, Harikrishna Narasimhan, and Oluwasanmi Koyejo. Fair performance metric elicitation. *arXiv preprint arXiv:2006.12732*, 2020.

[Ilv20] Christina Ilvento. Metric learning for individual fairness. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[JKN$^+$19] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *arXiv e-prints*, pages arXiv–1905, 2019.

[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.

[LMJ20] Lydia T Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pages 1618–1628. PMLR, 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.

[MP16] Debmalya Mandal and David C Parkes. Correlated voting. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 2016.

[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.

[SLB$^+$21] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

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

[ZYB21] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.