Learning Human Preferences for Socially Responsible AI

Debmalya Mandal

Max Planck Institute for Software Systems

Over the last decade, rapid development in deep learning methods has led to significant improvement in several areas of AI. Moreover, the success of deep reinforcement learning in large-scale games make us believe that these techniques can be used for making complex interactive decisions. Indeed the next frontier of AI lies in building reliable decision making systems that our society can utilize for making consequential decisions. We can envision a future where AI is used to build a financial plan for a country, or planning optimal allocation of scarce resources to citizens.

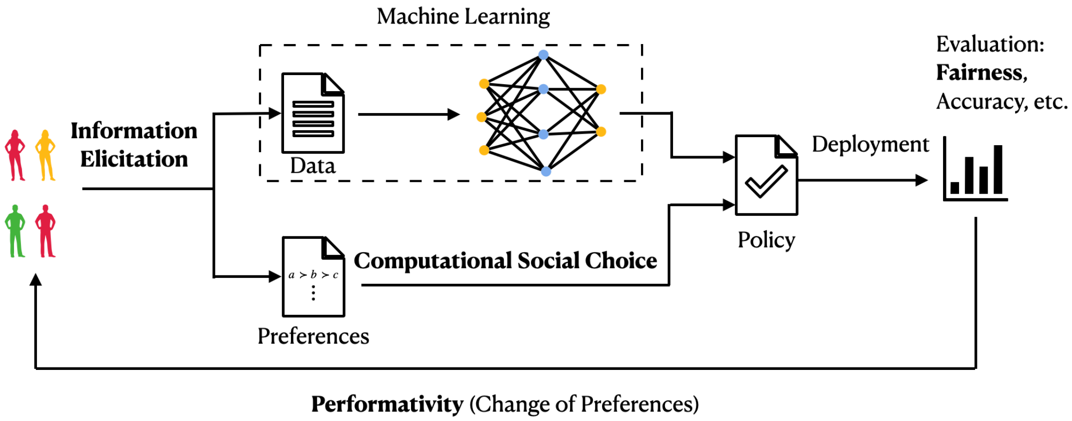
Machine Learning
Evaluation:
Fairness,
Accuracy, etc.
Information Elicitation
Data
Deployment
Policy
$a > b > c$
Computational Social Choice
Preferences
Performativity (Change of Preferences)

Figure 1: A Framework of AI for societal decision-making: Most of the recent developments focused on machine learning (shown in dotted box). For societal decision making, we need to learn human preferences, and understand performative effects of AI on society. My research focuses on developing methods (shown in bold) for addressing these challenges.

However, as figure 1 shows, machine learning is one component of the entire pipeline of a societal AI system. Standard learning algorithms assume existence of large datasets but they need to be collected from people through information elicitation mechanisms. Additionally, machine learning algorithms do not directly translate to an implementable policy and needs to understand people's preferences over various choices. These choices often involve concerns such as fairness and long-term impact of policies in society. Therefore, the success of AI for societal decision-making crucially depends on equal progress in the other components, and there are many challenges in these areas. I now highlight several challenges in developing AI systems for societal decision-making.

1. **Learning Human Preferences**: Any human-facing AI system needs to learn preferences of humans by interacting with them over time. This is particularly true for building public participation platforms where people have heterogeneous preferences. Additionally, this learning problem is often complicated as humans have inherent biases, and they also need to be incentivized to explore choices that are not necessarily popular.

2. **Performativity of Decision-Making Systems**: Even if human preferences can be learned for designing the right objective function, the decisions made by a system (e.g. a policy deployed by a reinforcement learning algorithm) can change the preferences of the users, and thereby the underlying environment where the system operates. Therefore, for societal decision making, we need to consider such long-term effects and revise our objective and training algorithms.

3. **Fairness in Multi-Agent Systems**: Most societal systems often have multiple stakeholders with different utility functions. For example, in resource allocation problems, there are different groups of users and selecting the wrong objective might introduce unfairness to the minority group. Additionally, the nature of the groups changes based on the application, and we must ensure that fairness of the multi-agent system is preserved when deployed on a new environment.

Computational Social Choice

- Communication-Distortion Trade-off [12, 13]
- Uncertain preferences [11]
- Surprisingly Popular Voting [4]

Algorithmic Fairness

- Robust optimization for fairness [7]
- Fairness in Online Learning [6, 8]
- Fair clustering [5]

Information Elicitation

- Heterogeneous Users / Tasks [1, 10]
- Long-Term Forecasting [9]

Sequential Decision Making

- Performative Reinforcement Learning [14]
- Learning with Blocking Constraints [2, 3]

Figure 2: Four main themes of my research interests

A significant part of my research is directed towards addressing the challenges of designing AI enabled societal decision-making systems highlighted above. In particular, I work on – *computational social choice* for learning and aggregating human preferences, *performative reinforcement learning* for modeling the impact of RL policies on the environment, and finally *algorithmic fairness* for designing of *fair* algorithms, particularly in multi-agent systems. Figure 2 summarizes my research in various areas, and next I describe them in detail.

1 Computational Social Choice

Social choice theory studies the design of voting rules in order to aggregate individual preferences into a collective preference. Moreover, the field of computational social choice is interested in settings with complex preferences over a large number of alternatives. This problem is of particular interest in platforms like *participatory budgeting* which decides how to allocate public budget by incorporating citizens' preferences over a large number of projects.

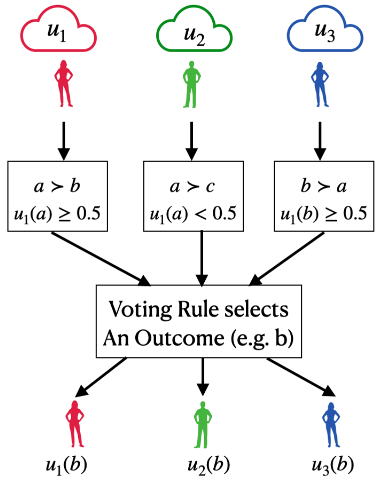

Figure 3: **Implicit Utilitarian Voting**: We introduce a new model of voting where voters can report general preferences (e.g. top k alternatives and additional cardinal information). This lets us study the trade-off between elicitation complexity and welfare of voting rules. [12, 13]

A drawback of existing voting protocols is that it only asks users to provide a rank over the set of possible alternatives. In a platform like the 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 papers [12, 13], 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 problem with a large number of alternatives is that existing voting rules often require many votes to converge to the correct answer. In a recent work [4], 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 outperforms the classical voting rules on questions from different domains.

2 Performative Reinforcement Learning

Recent success of deep reinforcement learning makes reinforcement learning a promising candidate for long-term societal decision making. Indeed several online platforms already use deep reinforcement learning for designing recommender systems. However, these applications often ignore the fact that the deployed policy changes peoples preferences and hence the underlying system over which the learner is optimizing. In order to capture such a phenomenon, we introduce the framework of performative reinforcement learning [14] where the policy chosen by the learner affects the underlying reward and transition dynamics of the environment. As figure 4 highlights the underlying Markov decision process is no longer fixed, but rather it is parameterized by the deployed policy.

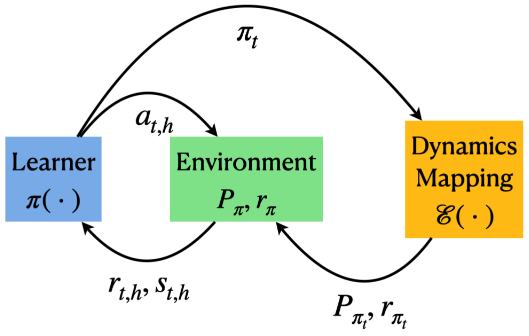

Figure 4: An illustration of the **Performative Reinforcement Learning** framework [14]. The underlying Markov decision process changes in response to the deployed policy. If the learner updates the policy π_t then the dynamics mapping $\mathcal{E}$ generates the new reward function r_{π_t} and transition function P_{π_t}.

One of the main goals of performative reinforcement learning is to find a stable policy, since such a stable policy can be deployed in a stable environment where the policy is also optimal. Our main contribution is to show that repeatedly optimizing a regularized version of standard reinforcement learning problem converges to a stable policy under reasonable assumptions on the transition dynamics. We then extend our results for the setting where the learner just performs gradient ascent steps instead of fully optimizing the objective, and for the setting where the learner has access to a finite number of trajectories from the changed environment. For both the settings, we leverage the dual formulation of performative reinforcement learning, and establish convergence to a stable solution.

Besides finding a stable policy, I am interested in two interesting questions in the realm of performative reinforcement learning. From the learner's perspective, finding a performatively optimal policy is more desirable as it guarantees that the policy performs optimally with respect to the changed environment. Second, from the perspective of society, it is important to find stable policies that don't polarize preferences, and don't drive away certain subpopulations from the platform. Solving this question will require careful modeling of the environment dynamics mapping $\mathscr{E}(\cdot)$ shown in figure 4.

3 Algorithmic Fairness

Most decision-making systems face diverse groups and the goal of fair decision making in AI is to ensure some measure of fairness across different groups. However, despite continued efforts to develop fair machine learning algorithms, there are several shortcomings in current systems and proposed definitions.

The literature on fair classification has largely ignored the design of fair and robust classifiers. In a recent work [7], 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 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.

The goal of fair reinforcement learning is to design long-term decision making system with multiple stakeholders. The main question is to come up with a fair objective that the system should optimize. In a recent work [8] we take an axiomatic view of this problem, and propose a set of axioms that such a fair objective must satisfy. We show that the Nash social welfare is the unique objective that satisfies all the axioms, whereas prior objectives like minimum welfare or generalized Gini welfare fail to satisfy all of them. We then consider the learning version of the problem where the underlying model i.e. Markov decision process is unknown. We consider the problem of minimizing regret with respect to the fair policies maximizing different fair objectives, propose a generic learning algorithm and derive its regret bound with respect to the different policies.

4 Other Research Interests

4.1 Information Elicitation

Most machine learning algorithms require large datasets for training and these datasets are obtained from real humans through crowdsourcing. Information elicitation studies the design of such mechanisms to incentivize people to provide accurate data. My research on *peer prediction* considers information elicitation for settings where the correctness of information cannot be verified, either because there is no objectively correct answer or because the answer is too costly to acquire.

Existing peer prediction mechanisms ignore the fact that users providing feedback may be quite different in the way they think about the world. Such heterogeneity is quite common in applications like grading peer assignments or reviewing services in online platforms. Our work [1] 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.

Peer prediction methods are also useful in improving accuracy of long-term forecasting. In a recent work [9], we demonstrated the effectiveness of peer-prediction based methods in long-term forecasting of geo-political 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.

4.2 Sequential Decision Making with Constraints

Multi-armed bandits are the standard models of sequential decision making under uncertainty. However, they usually ignore that services (i.e. arms) are often unavailable for some number of rounds once they are allocated (i.e. pulled). In a recent work [2], we model this problem through *adversarial blocking bandits* where the rewards and the blocking lengths of the arms can vary arbitrarily over time. We first show that the offline version of the problem i.e. known rewards and blocking lengths, is NP-hard and we construct a greedy policy that is a constant factor approximation of the optimal policy. For unknown reward functions, we design a learning algorithm that has sublinear regret when measured with respect to the greedy benchmark.

The blocking constraints are also prevalent in repeatedly matching multiple agents to multiple services. In such an online matching problem, the value of matching a user to a service is a priori unknown, and the system must learn how much value a user receives when assigned to a particular service. In a recent work [3], we apply the framework of *adversarial blocking bandits* to multi-agent matching problem where allocating a service to a user blocks that service for a 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 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 Future Research Directions

Here I briefly provide an overview of some of my future research interests. I am particularly interested in developing fair learning algorithms for multi-agent systems. As multi-agent systems become more wide-spread throughout our society, we need to ensure that different agents cooperate to solve a fair objective. Moreover, most of the learning agents operate independently, and a challenging question is whether we can develop decentralized learning algorithms for MARL that achieve a desired notion of fairness.

I am also interested in understanding and modeling the effects of reinforcement learning on human preferences. Performative reinforcement learning [14] proposes a general framework of how a policy changes human preferences, and also the underlying environment. However, in order for this framework to be usable in real-world systems like recommender systems, we need more careful modeling of the environment dynamics. Additionally, performative value functions are non-convex, so optimizing under changing environment dynamics is challenging.

References

[1] 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.

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

[3] Nicholas Bishop, Hau Chan, Debmalya Mandal, and Long Tran-Thanh. Sequential blocked matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4834–4842, 2022.

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

[5] Debajyoti Kar, Sourav Medya, Debmalya Mandal, Arlei Silva, Palash Dey, and Swagato Sanyal. Feature-based individual fairness in k-clustering. *arXiv preprint arXiv:2109.04554*, 2021.

[6] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[7] 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.

[8] Debmalya Mandal and Jiarui Gan. Socially fair reinforcement learning. *arXiv preprint arXiv:2208.12584*, 2022.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. *arXiv preprint arXiv:2207.00046*, 2022.