AI for Societal Decision Making

Debmalya Mandal

Max Planck Institute for Software Systems

Over the last decade, rapid development in machine learning techniques has led to significant advances in AI capabilities. In particular, the success of reinforcement learning in domains ranging from chemistry [18] to robotics [2], suggests that these techniques can be applied to support complex interactive decisions for society. The next frontier of AI lies in building reliable decision making systems that society can utilize in supporting consequential decisions, for example building a financial plan for a country, allocating resources such as vaccines to citizens, or even regulating a complex financial market.

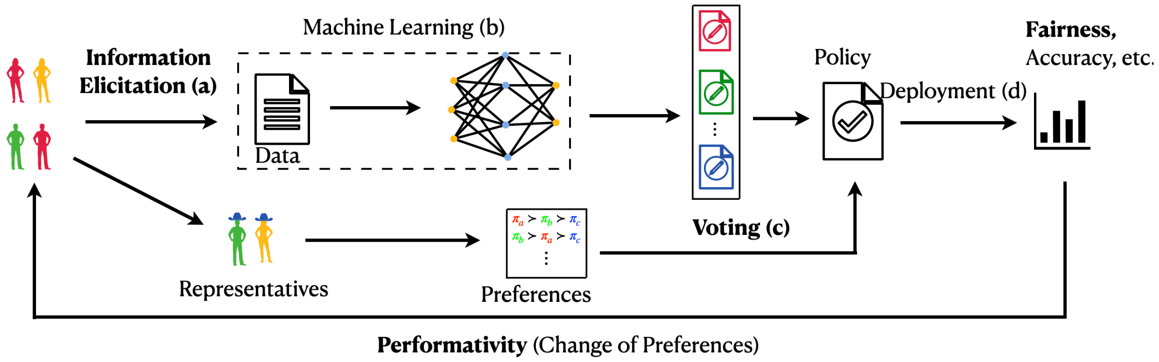
Information Elicitation (a)
Machine Learning (b)
Data
Fairness,
Accuracy, etc.
Policy
Deployment (d)
Voting (c)
Representatives
$\pi_a \succ \pi_b \succ \pi_c$
$\pi_b \succ \pi_a \succ \pi_c$
$\vdots$
Preferences
Performativity (Change of Preferences)

Figure 1: A Framework of AI for Societal Decision-Making: Most of the recent developments in AI have focused on machine learning (shown in the dotted box) and deployment (steps b and d). For effective societal decision making, we need to incorporate human preferences over policy decisions, and understand performative effects of AI on society. My research focuses on developing methods (shown in bold) for addressing these challenges.

However, in order to utilize AI for societal decision-making, we need to rethink the entire machine learning pipeline from data elicitation to deployment. In particular, any such framework must involve human preferences for selecting policies, and consider the long-term effects of policies in our society. Figure 1 presents such a framework with four main steps – (a) **information elicitation** (collect data about human behavior and build a model e.g. through imitation learning), (b) **machine learning** (use a learning algorithm to propose a set of models), (c) **voting** (use a voting rule to choose a policy based on representatives' preferences), and (d) **deployment** (evaluate model performance through metrics).

Most of the recent developments in AI have focused on machine learning and deployment. However, as Figure 1 shows, machine learning is but one component of the entire pipeline of an AI system. The success of AI for societal decision-making will crucially depend on equal progress being made in the other components, and there remain many challenges:

(A) **Aggregating Human Preferences**: Reinforcement learning promises to be an exciting tool for designing complex economic policies [20]. However, such an AI-generated policy must be validated by experts before deployment. This can be achieved by proposing a set of policies to the experts and then using a voting rule to aggregate their opinions to select a final policy. However, there are two challenges – comparing two policies (e.g. tax policies) might not be easy even for the experts since the long-term impact of a policy is not always obvious. Moreover, the range of policies a learning algorithm can produce is huge.Therefore, an important problem is to elicit and aggregate experts' opinions over a large space of complex policies.

(B) **Fairness in Multi-Agent Systems**: Recommender systems often face a diverse group of users with different utility functions. Fairness in such online systems is an important concern and a major challenge is to select an objective for the platform that provides desired fairness guarantees across different groups. Furthermore, the population faced by a recommender system is dynamic, and uncertain which means that the decomposition of the population into different groups is not known a priori. Therefore, a challenging problem is to design a fair learning algorithm that is robust to such variability of the groups within the population.

(C) **Performativity of Decision-Making Systems**: Most online platforms interact with the users over multiple rounds, and use reinforcement learning for recommendation. However, such recommender systems can change how frequently the users interact with the platform, and also their underlying true preferences. This implies that reinforcement learning based recommender systems are *performative* in the sense that the deployed policy can change the underlying environment i.e. Markov decision process. These effects are often not immediate and visible only over a period of time. Therefore, we need to consider such performative effects of learning algorithms and reformulate our objective that promotes improved long-term welfare of the population.

Computational Social Choice

- Communication-Distortion Trade-off [13, 16]
- Uncertain preferences [12]
- Surprisingly Popular Voting [6]

Sequential Decision Making

- Performative Reinforcement Learning [17]
- Online Learning with Constraints [3, 4]
- General Models of Discounting in RL [5, 14]

Algorithmic Fairness

- Robust optimization for fairness [9]
- Fairness in Online Learning [8, 10]

Information Elicitation

- Heterogeneous Users / Tasks [1, 11]
- Long-Term Forecasting [15]

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 (see Figure 2). In particular, I work on – *computational social choice* for eliciting and aggregating human preferences, *performative reinforcement learning* for modeling the impact of RL on society, and *algorithmic fairness* for designing *fair* learning algorithms.

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 particularly relevant for validating AI-generated policies by human experts. As discussed in challenge (A), the range of possibilities for an economic policy such as a redistribution mechanism can be huge, and comparing two such policies can be difficult for an expert. Therefore, an open question is to design appropriate voting rules on the space of complex policies. For a simple abstraction, consider the aggregation of n experts' preferences over m (possibly large compared to n) alternatives, in order to select a single alternative.

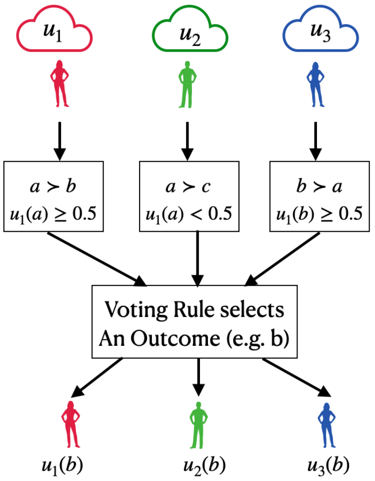

Figure 3: A new model of voting where voters can report general preferences. This lets us achieve optimal trade-off between elicitation complexity and social welfare of voting rules. [13, 16]

Traditional voting rules ask voters to provide a rank order on the set of possible alternatives. This has poor performance when there is a large number of alternatives because a ranking doesn't reveal any information about the relative strength of the voters' preferences. In its place, I consider asking users to reveal additional cardinal information. In a pair of papers [13, 16], I have characterized the trade-off between the achievable, utilitarian social welfare and the communication complexity of voting rules, which measures the amount of information the voters must convey to the voting rule. For the upper bound, I propose new voting rules that use sketching algorithms for the first time. The lower bound makes interesting connections between the literature on communication complexity and voting.

Another problem with classical voting rules is that they can fail when the experts with the correct opinion are the minority group in the population. In a recent work [6], I show how to alleviate this problem by asking voters additional prediction questions about the opinions of others. A large-scale experiment on Amazon Mechanical Turk shows that the combination of votes and prediction reports outperforms classical voting rules on questions from different domains.

2 Algorithmic Fairness

In recent years, researchers have invested significant efforts in designing fair learning algorithms, ranging from de-biasing training datasets to designing algorithms with explicit fairness constraints. However, it remains challenging to design a fair classifier that is robust to variability in the dataset. Many datasets are skewed in terms of the demographics of labellers and the representation of minority groups. Furthermore, as discussed in challenge (B), in many domains including healthcare, different subpopulations have different stakes over the decisions of the learning algorithm, and the designer needs to address these trade-offs. In my research, I have used principles from economics, and fair division in particular, to address this challenge.

Many "fair" classifiers are actually sensitive to the representation of different groups in the training data. In fact, I show that existing classifiers become grossly unfair even if we slightly perturb the training distribution [9]. Since test distribution is often different than the training distribution, the fairness guarantees of these classifiers do not carry over to deployment. I study classifiers that are fair not only with respect to the training distribution, but also for a class of distributions that are perturbed from the training data. I set up training with a min-max objective function, whose goal is to find a classifier that is fair with respect to a large class of distributions. Experiments on standard benchmarks 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.

In the context of fair reinforcement learning, where the goal is to deploy a policy that doesn't discriminate against any subpopulation, a concern is to select the right objective for optimization. In a recent work [10], I take an axiomatic view of this problem, and propose a set of axioms that such a fair objective must satisfy. I show that the Nash social welfare is the unique objective that satisfies all the axioms, and also consider the learning version of the problem, where the underlying model is unknown. I propose a generic learning algorithm for minimizing regret with respect to different fair policies and bound its regret.

3 Performative Reinforcement Learning

Existing frameworks for reinforcement learning often ignore the fact that the deployed policy might change the underlying environment (reward, transition probability, or both). For example, recommender systems often use contextual Markov decision processes to model the interaction with a user. In a contextual MDP, the initial context/user feature is drawn according to a distribution, then the user interacts with the platform according to the context-specific MDP. However, it has been observed that these systems not only change the user demographics (i.e. distribution of contexts) but also how they interact with the platform. In order to capture such a phenomenon, I introduced the framework of *performative reinforcement learning* [17], where the policy chosen by the learner affects the underlying reward and transition dynamics (see Figure 4).

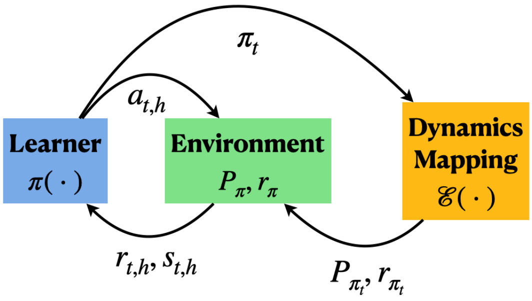

Figure 4: The framework of **Performative Reinforcement Learning** framework [17]. The underlying MDP changes in response to the deployed policy. If the learner updates the policy π_t then the dynamics mapping $\mathcal{E}$ generates the new reward function r_{π_t} and transition function P_{π_t}.

One of the main goals of performative reinforcement learning is to find a stable policy π_S, which is optimal with respect to the changed MDP $M(\pi_S)$. Our main contribution is to show that repeatedly optimizing a regularized version of a standard reinforcement learning problem converges to a stable policy under reasonable assumptions on the transition dynamics.

We also extend our results for the setting where the learner only has access to a finite number of trajectories from the changing environment. Our proof leverages an interesting dual characterization of performative RL and introduces a new perspective in analyzing the convergence of algorithms with decision-dependent environments.

Besides finding a stable policy, I am currently working on several interesting questions in performative RL. First, extending our result [17] to a general function approximation setting is quite interesting. Note that, it is not even clear when a stable policy exists for with general function approximation. Second, from the learner's perspective, finding a performatively optimal policy might be desirable if one cares about immediate reward with respect to the changed MDP.

Finally, it would be interesting to consider specific models of the dynamics mapping ($\mathscr{E}(\cdot)$ in figure 4). For example, it is known that exploration in online reinforcement learning directly affects how long the users stay in the platform or equivalently, how they discount their future utilities. Compared to the standard geometric discounting, RL with general discounting is relatively less studied and has been the focus of some of my recent works [5, 14]. It would be interesting to see what happens if the discounting factor is affected by the policy, and how to minimize regret in such a performative setting.

4 Other Research Interests

4.1 Information Elicitation

Large datasets for machine learning are often created through crowdsourcing. The CIFAR-10 dataset, for example, was created by paying students to label data. Information elicitation studies the problem of designing such mechanisms to obtain accurate data from people. My research on peer prediction looks to design information elicitation schemes for settings where the correctness of information cannot be verified. I have designed incentive-aligned peer prediction methods when the participating users are heterogeneous [1] as well as for heterogeneous tasks [11]. I have also demonstrated the effectiveness of peer-prediction in long-term forecasting of geo-political events, including climate change, and sports [15].

4.2 Sequential Decision Making with Constraints

Online matching considers the problem of repeatedly matching services to users over time, under the uncertainty of the reward obtained from a matching. However, one of the difficulties in repeated matching is blocking — a service gets blocked for a variable number of rounds once allocated to a user. This problem arises, for example, in matching renewable energy consumers to producers in European energy markets. In a recent work [3], I introduced the adversarial blocking bandits model to study the single-agent version of this problem, and derived learning algorithms for this setting. Subsequently, I have applied this framework to the problem of online matching with blocking constraints [4]. This work was the first to derive a multi-agent learning algorithm with per-agent logarithmic regret for the problem with blocking constraints.

5 Future Research Directions

5.1 Fairness in Multi-Agent Systems

As multi-agent systems become more wide-spread in settings such as autonomous driving, and online platforms, we need to promote cooperation between agents and ensure that they work together towards a fair objective. I am particularly interested in designing decentralized learning algorithms for achieving a joint policy satisfying desired fairness guarantees for the agents. As an example, consider the problem of designing negotiation protocols among different countries for tackling climate change. Recently countries like Switzerland and Germany have started investing in clean energy infrastructures in poorer nations like Georgia, and Ghana in exchange of CO_2 emission credits [19]. Such a bilateral agreement is beneficial for both the countries in short-term, but might not be optimal or fair in the long-run. The selection of a fair and optimal negotiation policy can be formalized as a multi-objective RL problem. In particular, each policy specifies a point in the GDP vs CO_2 emission plot for the two participating countries, and a first goal should be to select a policy that is not pareto-dominated by any other policy.

My recent work on fair reinforcement learning [10] considers a setting where n different agents have n different reward functions, and the goal is to choose a policy that maximizes a fair objective (e.g. Nash social welfare). The problem of bilateral negotiation can be generalized to n different agents with m different metrics. It would be interesting to design learning algorithms that converge to pareto optimal policies for the general problem. For the decetralized version of this problem, the main challenge is reducing communication among the agents which could be huge with a large number of agents. Additionally, there are privacy issues with communicating sensitive data, particularly for designing negotiation protocols between different nations. From a theoretical point of view, it would be interesting to see how much communication is necessary to attain regret that is sub-linear in the horizon length.

5.2 Population Dynamics under Reinforcement Learning

Data driven decision making systems are ubiquitous in our society, ranging from recommender systems to online matching platforms. Although these systems use RL to learn about user preferences they are mostly designed to maximize user engagement with the platforms and numerous studies have shown that they have adverse effects on users' preferences. For example, randomized controlled trials have shown that platforms like Facebook tend to expose people to political news matching their ideologies [7], leading to polarized opinions over time. Therefore, a major challenging question is whether we could use RL in such a way that these systems satisfy certain desirable behaviors over long-term.

As an example of undesirable effects of data driven decision making, consider the problem of opinion polarization in recommender systems. Recently proposed geometric model of opinion dynamics provide some understanding behind the phenomenon of opinion polarization. This model assumes that a user's preference changes in alignment with the recommended item. The general problem of recommending items can be formulated as an RL problem where the state space is the opinions of the users. Then the main question is whether we can design reward function so that the optimal policy doesn't lead to polarized opinions. Moving beyond the geometric model, contextual MDPs are often used in recommender systems. My recent work on performative RL [17] shows how repeatedly optimizing a regularized objective converges to a stable policy in these models. However, we would like to control either the optimization method or the dynamics, so that the resulting limiting policy doesn't cause polarization of preferences.

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*. 2017, pp. 81–98.

[2] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. "Learning dexterous in-hand manipulation". In: *The International Journal of Robotics Research* 39.1 (2020), pp. 3–20.

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

[4] Nicholas Bishop, Hau Chan, **Debmalya Mandal**, and Long Tran-Thanh. "Sequential blocked matching". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 5. 2022, pp. 4834–4842.

[5] Jiarui Gan, Annika Hennes, Rupak Majumdar, **Debmalya Mandal**, and Goran Radanovic. "Markov Decision Processes with Time-Varying Geometric Discounting". In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence* (2023).

[6] Hadi Hosseini, **Debmalya Mandal**, Nisarg Shah, and Kevin Shi. "Surprisingly Popular Voting Recovers Rankings, Surprisingly!" In: *The Thirtieth International Joint Conference on Artificial Intelligence* (2021).

[7] Ro'ee Levy. "Social media, news consumption, and polarization: Evidence from a field experiment". In: *American economic review* 111.3 (2021), pp. 831–70.

[8] Yang Liu, Goran Radanovic, Christos Dimitrakakis, **Debmalya Mandal**, and David C Parkes. “Calibrated fairness in bandits”. In: *Fairness, Accountability, and Transparency in Machine Learning* (2017).

[9] **Debmalya Mandal**, Samuel Deng, Suman Jana, Jeannette M Wing, and Daniel Hsu. “Ensuring Fairness Beyond the Training Data”. In: *Advances In Neural Information Processing Systems* (2020).

[10] **Debmalya Mandal** and Jiarui Gan. “Socially Fair Reinforcement Learning”. In: *arXiv preprint arXiv:2208.12584* (2022).

[11] **Debmalya Mandal**, Matthew Leifer, David C Parkes, Galen Pickard, and Victor Shnayder. “Peer Prediction with Heterogeneous Tasks”. In: *NIPS 2017 Workshop on Crowdsourcing and Machine Learning* (2017).

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

[13] **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*. 2019, pp. 7180–7191.

[14] **Debmalya Mandal**, Goran Radanovic, Jiarui Gan, Adish Singla, and Rupak Majumdar. “Online Reinforcement Learning with Uncertain Episode Lengths”. In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence* (2023).

[15] **Debmalya Mandal**, Goran Radanovic, and David C Parkes. “The Effectiveness of Peer Prediction in Long-Term Forecasting.” In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence*. 2020, pp. 2160–2167.

[16] **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*. 2020, pp. 795–813.

[17] **Debmalya Mandal**, Stelios Triantafyllou, and Goran Radanovic. “Performative Reinforcement Learning”. In: *arXiv preprint arXiv:2207.00046* (2022).

[18] Marwin HS Segler, Mike Preuss, and Mark P Waller. “Planning chemical syntheses with deep neural networks and symbolic AI”. In: *Nature* 555.7698 (2018), pp. 604–610.

[19] Hiroko Tabuchi. “Switzerland Is Paying Poorer Nations to Cut Emissions on Its Behalf”. In: *The New York Times* (Nov. 7, 2022).

[20] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher. “The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning”. In: *Science advances* 8.18 (2022), eabk2607.