Extracting Incentives From Black-Box Decisions

Yonadav Shavit*[1], William S. Moses*[2]

Harvard SEAS [1], MIT CSAIL [2]

yonadav@g.harvard.edu [1], wmoses@mit.edu [2]

High-Level Overview

Every algorithmic decision-maker incentivizes people to act in certain ways to receive better decisions. These incentives can dramatically influence subjects' behaviors and lives, and it is important that both decision-makers and decision-recipients have clarity on which actions are incentivized by the chosen model.

Why the incentives provided by algorithms matter:

- They are **legally regulated** (e.g. adverse action notices in credit scoring).
- They **empower individuals** to have control and agency over their own outcomes (e.g. [2]).
- Whether we study them or not, **all algorithms already incentivize behaviors**, and are having unobserved consequences for decision-makers and decision-recipients in the real world.

In this work, we propose a novel framework for analyzing algorithmic incentives through the lens of **Markov decision processes (MDPs)**.

At a high level, we propose that to properly understand how an individual is incentivized to act, we must first define the actions available to an individual, and their effects. Then, the individual is **incentivized** to take whichever action will modify their current state such that, after executing a sequence of additional actions, they will reach a final state that **maximizes their received decision**.

We show, using this framework, that many **traditional interpretability tools** (e.g. LIME[1], input gradients) can **provide poor advice policies** when the decision-making model is non-linear.

Our key contribution is **a method for identifying approximately optimal algorithmic incentives**, by using planning algorithms like MCTS to solve the agency MDP. Furthermore, our method is model-independent and requires only query access to the decision-making model.

We show in experiments that this method outperforms local approximations' advice in practical settings, including an **online FICO scoring API**, and a random-forest-based **violent recidivism predictor**.

Framework

Consider an individual $s \in \mathcal{S}$, defined as a feature vector. Individual s wants to maximize the outcome of positive-definite decision function $D(s) \in \mathbb{R}^+$.

By taking an action $a \in \mathcal{A}$, individual s may change their state. Individual s's next state is defined by sampling from transition model $\mathcal{T}$ where $s' \sim \mathcal{T}(s, a)$.

We combine $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{T}$ together to form a Markov Decision Process (MDP). We specify a terminal function end(s) that determines whether the sequence has ended, and define the reward function $\mathcal{R}$:

$$\mathcal{R}(s) = \begin{cases} D(s) & \text{if end}(s) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

An **advice policy** $\pi \in \Pi$ recommends a certain action $a = \pi(s)$ for each state s an individual may encounter.

For example, a locally-optimizing **greedy policy** chooses actions based only on maximizing the immediate improvement in the received decision:

$$\pi_{local}(s) = \arg\max_{a \in \mathcal{A}} \mathop{\mathbb{E}}_{s' \sim \mathcal{T}(s,a)} [D(s')] \tag{2}$$

We say an action is **incentivized** if it is recommended by an **optimal advice policy** π^*. More specifically, an individual with state s is incentivized to execute action a^* if that action will maximally improve their eventual expected decision, more than any alternative action:

$$a^* = \max_{\pi \in \Pi} \left(\mathop{\mathbb{E}}_{s_{final} \sim H_\pi(s)} [D(s_{final})] \right) = \pi^*(s) \tag{3}$$

where $H_\pi(s)$ is the distribution of end-states resulting from "rolling out" π starting at state s.

We can approximate this optimal advice policy by leveraging **planning algorithms such as reinforcement learning**.

Problems with Local Approximations as Advice Policies

Local-approximation-based advice can be dangerously wrong. Consider the example in Figure 1, in which a locally-improving policy would trap the individual at a local maximum and never achieve the better outcome that was available to them. Moreover, local advice may still be sub-optimal when the decision function is monotonic as is the case in Figure 2.

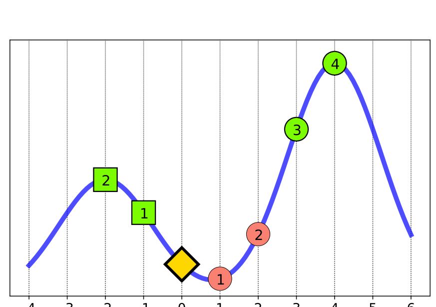

Figure 1. A subject (gold diamond) wants to maximize the value from a decision after moving at most 4 units from the origin. If the individual has two resources or fewer, they should head towards the left, whereas if they have 3 resources or more, they should move right.

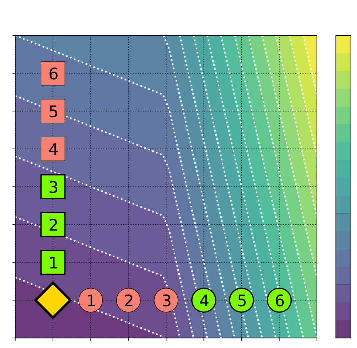

Figure 2. A 2D monotonic decision function, with the lowest output in bottom-left (blue). A subject starting in the bottom-left corner can move 1 grid unit each step. Greedy is optimal given at most 3 resources, but misses the optimal policy for 4 or more steps.

Experiments

We applied our incentive-evaluation framework to two decision-settings: **pretrial risk assessment** (based on the COMPAS dataset), and **credit scoring** (by querying FICO's online credit score calculator).

We also trained a double deep Q-network on the agency MDP, but found that in both settings the network generally failed to learn a meaningful advice policy (not even equaling the greedy policy), and so we have excluded those results.

We compare these incentives to the **greedy** policy (Eq. 2), which maximizes the decision immediately after the current action, and to a **random** policy.

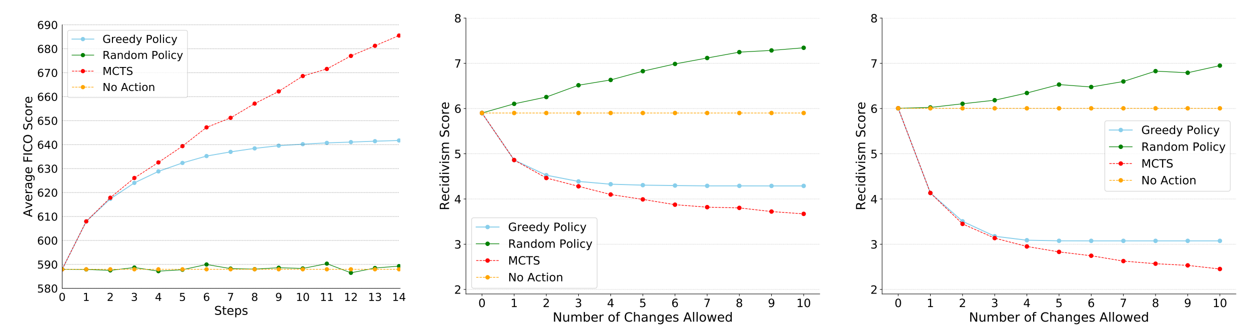

Figure 3. Comparing the performance of different advice policies, as defined in Eq. 3, varying the initial resource count. **Left:** Simple Credit model (averaged over 1000 initial states, higher is better). **Center:** recidivism prediction, including race and gender (averaged over 1000 initially medium/high-risk states, lower is better). **Right:** Recidivism prediction (excluding race/gender).

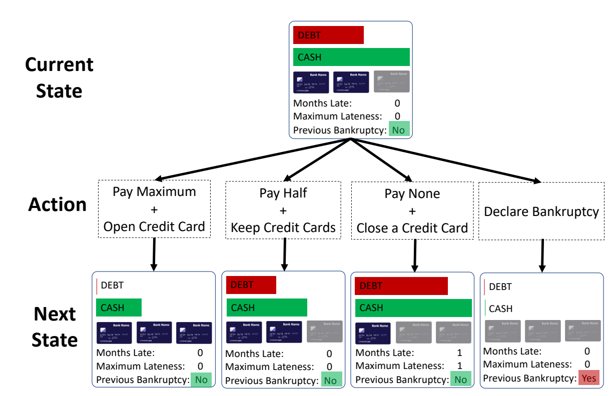

Figure 4. Examples of actions an agent can take each month within the "complex FICO" MDP.

We also tested our framework in a more complicated credit scoring setting, with realistic actions that each affect multiple features. For examples of some of the actions, see Figure 4.

We can see the effect of different advice policies on loan recipients in different financial scenarios in Figures 6 and 7.

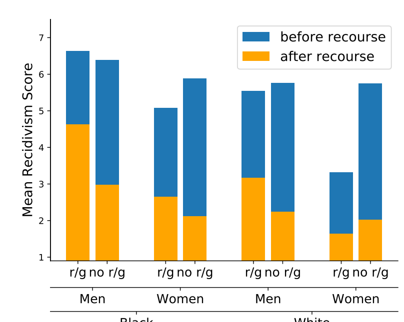

Figure 5. Mean recidivism risk score before and after following MCTS-generated incentives for 6 to 10 steps, varying the inclusion of race/gender in decisions.

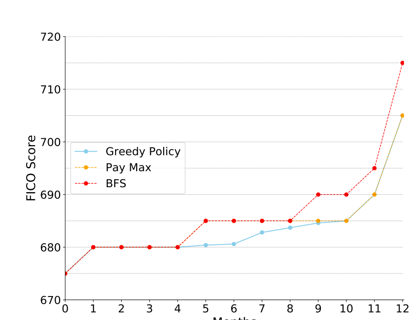

Figure 6. Credit score under a realistic model, starting with US average financial data and no debt, and varying time remaining before score is checked.

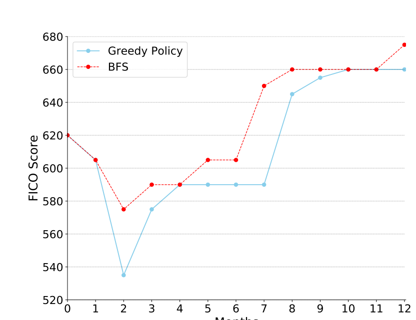

Figure 7. Credit score under a realistic model, starting with US average financial data but a sudden crisis of $10,000 of debt, and varying time remaining.

Conclusions

- Incentives and agency are crucial concepts to study further, and are almost certainly impacting decision-subjects' behavior in unknown and unfair ways.
- Our method can successfully learn underlying incentives even from black-box APIs, and from realistic action spaces.
- Local linear approximations fail and provide suboptimal advice in real-world models, and the use of such models today is misleading both decision-recipients and decision-makers.
- Many open questions remain, including how well different interpretability schemes generate advice, and how to efficiently construct action models in the real world
- **We'd love to talk about collaborating to identify incentives in real-world systems!**

Acknowledgements & References

We would like to thank Christina Ilvento for her early insights into the challenges of identifying incentives, and to Cynthia Dwork, Michael P. Kim, Suhas Vijaykumar, Boriana Gjura, and Charles Leiserson for their helpful feedback and advice.

William S. Moses was supported in part by a DOE Computational Sciences Graduate Fellowship DE-SC0019323. This research was supported in part by NSF Grant 1533644 and 1533644, in part by LANL grant 531711, and in part by IBM grant W1771646.

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.
Why should i trust you?: Explaining the predictions of any classifier.
In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

[2] Berk Ustun, Alexander Spangher, and Yang Liu.
Actionable recourse in linear classification.
In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19. ACM, 2019.