Etan Green | Research Statement | December 2021

In recent years, social scientists have begun applying tools from machine learning to one-shot policy problems—in which the policy maker makes a single decision—such as whether to detain a defendant before trial. However, many important policy problems—such as setting interest rates or prescribing medicine to patients—are *dynamic*, meaning that a sequence of decisions determines the eventual outcome. Reinforcement learning (RL) is a powerful tool for solving dynamic problems. So far, RL has been used mainly to play canonical games such as chess and Go or to train robots to mimic humans. I use RL to solve real-world, dynamic policy problems.

One such problem is bargaining. Bargaining is ubiquitous in markets, but the massive academic literature on bargaining—which spans economics, psychology, and management—has no answer to what is arguably the most fundamental question: what offer is best? This question is dynamic in nature: in sequential bargaining, each party may make multiple offers before the negotiation concludes. In "Optimal Bargaining on eBay Using Reinforcement Learning," my coauthor (a former Penn undergraduate) and I train an RL agent to bargain optimally against humans in Best Offer listings on eBay, perhaps the largest bargaining market in the world.

The paper's first contribution is to characterize optimal bargaining on eBay in a manner that humans can use to bargain better. As a buyer, the RL agent bargains more aggressively than humans. The agent makes lower first offers than human buyers, and it makes more offers—and it buys more items for less money as a result. As a seller, the RL agent reads signals that human sellers ignore, and it sells more items for more money as a result. Whereas human sellers accept high first offers, for instance, the agent seller rejects them because high first offers signal buyers' willingness to pay more. For both the buyer and seller, we present simple strategies that allow human sellers to approach the performance of our deep RL agents.

The paper's second contribution is a new method for training RL agents to solve real-world dynamic policy problems. In bargaining, as in any real-world domain, training RL agents online (i.e., in the real world) is logistically infeasible: RL algorithms learn a good policy only after millions of actions. We pioneer a different approach: training a simulator from data, and training RL agents in that simulator. Specifically, we train a set of neural networks to mimic the behavior of humans as observed in a very large dataset of eBay listings, and we then train RL agents to play optimally against these simulated humans. This approach—training simulators from data—has the potential to dramatically expand the set of policy problems that RL can address.

One such problem is pricing. Algorithmic attempts at price setting often simplify the problem by treating it as a one-shot game: what price maximizes profit now? This myopia ignores at least two future concerns. Discounts cannibalize future sales by 1) inducing customers to stock up and 2) training customers to expect lower prices. Neglecting these concerns likely leads to over-discounting.

I am working with a global business-to-business e-commerce platform to set prices for products sold through their app. We treat pricing as a reinforcement learning problem: what prices optimize a discounted stream of profits over a long horizon? My work on bargaining provides a useful template. We first use the site's clickstream data to train a simulated customer, who responds to a given price schedule with a vector of purchase decisions. We then train an RL agent, which chooses a price schedule to optimize purchase decisions by the simulated customers over time.

When setting prices on the app, the agent strikes a balance between following what it believes to be the optimal policy and exploring other policies. This allows us to address a key limitation of the eBay project—that the agents can only choose among policies observed in the data. Online exploration allows the next version of the agent to search the policy space more widely during offline training. Important research questions that are uniquely answerable in this setup include: 1) How much is lost by treating pricing as a static problem, and why? And 2) How much is gained by setting different prices for different customers at the same time?

A dynamic problem of interest to game theorists, and to millions of game enthusiasts, is how to play poker optimally. In recent years, much has been written about how computers have beaten the best humans at poker. Less has been written about the brittleness of these algorithms: they are trained to play a version of Texas Hold'em in which each player begins every hand with a fixed stack (i.e., number of chips). To learn a different configuration, these algorithms have to start from scratch.

I am training a poker-playing agent using a simple and more robust approach: self-play reinforcement learning, in which the agent learns by playing against itself. So far, the agent performs nearly as well as leading computer poker players in the rigid setting in which leading agents are trained. However, the agent also performs well in settings where leading agents falter: when playing Greek Hold'em, a variant of Texas Hold'em, and when playing hands in which the players begin with different stacks. This approach has the potential to democratize computer poker. It requires an order of magnitude fewer computer resources to train than leading algorithms, and its implementation will be the first to be open-source.

Each of these projects, as well as prior work that does not use RL, shares a common theme: what does optimal behavior look like in complex decision-making domains? And how does human behavior compare? Two other papers explore variations on this theme.

"Bayesian Instinct" analyzes the decisions of home plate umpires in Major League Baseball. Optimal umpire behavior is clear: umpires should call a strike if the pitch intersects the official strike zone and a ball otherwise. Only the location of the pitch should determine the umpire's call. In practice, umpires enforce different strike zones in different counts, or game states.

Previous research labels this behavior a bias. We show that umpires are not biased—they are Bayesian. If the umpire is unsure whether the pitch is a ball or a strike, he should lean on his prior beliefs about where the pitch was likely to be thrown. We show that rational priors vary significantly across counts. Hence, a Bayesian umpire with imperfect vision will make different calls in different counts for pitches at the same location.

The paper uses a variety of tools to show that a Bayesian model accurately predicts the enforced strike zone in every count. Rational prior beliefs are estimated non-parametrically. Model parameters are estimated structurally. And the enforced strike zone is estimated using a support vector machine, a powerful supervised learning algorithm that finds an optimal boundary between locations where umpires call balls from locations where they call strikes.

"The Favorite-Longshot Midas" examines a market with a clear equilibrium prediction. In horse-race parimutuel markets, expected returns should be the same for all wagers. In practice, wagers on favorites, or horses that are more likely to win, return more in expectation than wagers on longshots. A long literature has debated explanations for this favorite-longshot bias, from risk-loving preferences to beliefs that overweight small probabilities.

We show the favorite-longshot bias instead results from deception by the track. In a parimutuel, the track makes money by taking a cut of the wagering; hence, its incentive is to increase wagering volume. Tracks make forecasts that systematically overweight the chances of longshots. This induces naive bettors to overbet longshots, and in doing so, creates an arbitrage opportunity for sophisticated bettors to wager on favorites. Were the forecasts accurate, naive bettors would correctly price the wagers, sophisticated bettors would not wager, and tracks would make less money.

The paper first documents a new stylized fact: the extent of the favorite-longshot bias varies across tracks. We then show, using a novel dataset containing the tracks' predictions, that the bias is only present at tracks which embed a favorite-longshot bias in their predictions. We write down a model that, using the track's predictions, predicts the observed variation in the bias across tracks.

Two earlier papers also use field data to explore interesting behaviors by experienced individuals. In "Personal Bests as Reference Points," my coauthor and I analyze 133 million online chess matches, finding that chess players care deeply about achieving their best-ever ratings. In a "Sharp Test of the Portability of Expertise," my coauthors and I find that professional basketball prognosticators with no statistical training can solve a difficult statistics problem—but only when formulated as a basketball prediction task rather than a math problem. We conclude that domain experience, rather than statistical knowledge, underpins the professionals' adeptness at statistical reasoning.