Evolving Sustainable Institutions in Agent-Based Simulations with Learning

Christopher Zosh* Andreas Pape† Todd Guilfoos‡ Peter DiCola§

August 4, 2025

Abstract

We develop a novel, game-theoretic computational model in which learning agents explore how much to consume from a common resource.

These agents live under three different political regimes: private provision, a benevolent and powerful social planner, and competitive direct democracy over vectors of (Pigovian) fines. Both agent consumption and voting decisions are guided by a single process: reinforcement learning with action similarity. The model produces panel data of fine vectors for each regime and setting.

We find the benevolent social planner's fines have significant welfare gains over uncoordinated private action, and that competitive direct democracy's fines can nearly achieve the same gains. We also find that learning changes the optimal solution: that is, the fine vector found by the benevolent social planner is both distinct from and performs better than the socially optimal fine vector analytically derived from this setting, were it populated with rational, fully-informed agents.

Elinor Ostrom empirically identified eight so-called "design principles" common to social structures of communities which successfully cultivate a common resource. One of these principles is "graduated sanctions," in which punishment accumulates at an accelerating rate as the degree of offense increases. We find that agent similarity is a necessary component for the emergence of graduated sanctions. We also find that, if fines generate revenue which can be costlessly redistributed, draconian (not graduated) sanctions emerge.

*Department of Economics, Binghamton University, United States of America

†Department of Economics, Binghamton University, United States of America

‡Department of Environmental and Natural Resource Economics, Rhode Island University, United States of America

§Pritzker School of Law, Northwestern University, United States of America

Keywords: Common Resources, Graduated Sanctions, Ostrom Design Principles, Learning, Agent-based Modeling

JEL Codes: D02, C63, P48, D83, D04, Q20

1 Introduction

We develop a computational model in which learning agents experiment with strategies in a public goods game and participate in forming bottom-up policy in an attempt to solve the commons problem. We pay special attention to graduated sanctions—sanctions that are initially small, but increasing and convex in the size of the infraction—to see whether and how this principle might emerge from the top down or bottom up in various contexts. We approach the problem iteratively by evaluating our model under three different political regimes and observing behavior during each. Under the Private Provision Regime, learning agents play a common-pool resource (CPR) game facing no policy. Under the Social Planner Regime, the learning agents play the same CPR game facing exogenously given policy via a benevolent social planner, who is experimenting with policies to improve social welfare. Finally, under the Competitive Direct Democracy Regime, the learning agents play the CPR game once more, but now participate in forming the policy they'll face by voting in a democracy. With these three models, we can explore behavior and outcomes when there is no policy, when policy is instituted from the top down, and when policy is formed by agent participation from the bottom up.

We vary a number of parameters to understand the institutional emergence of graduated sanctions. We vary how agents learn, the redistribution of taxes from punishment schemes, and democratic institutions. We use the results to interpret the sufficient conditions for graduated sanctions to emerge in a social dilemma. Our framework allows for us to investigate institutional factors that cannot be plausibly explored in traditional experiments or field studies.

We find that graduated sanctions emerge when a top-down social planner utilizes fines without redistribution, but only when agents utilize similarity in their decision making. When policy makers redistribute fines, however, draconian-style sanctions emerge instead as a more effective method of maximizing social welfare. We also find that when agents participate in a bottom-up policy selection by voting, they are able to solve the commons

problem nearly as well as the social planner. We also delve deeper into why both the theoretical solution and democracy achieve suboptimal levels of social welfare.

We use these agent-based models to build on previous results from the experimental literature. We show that specific mechanisms are important to the emergence of graduated sanctions, such as the concept of similarity of behavior. Similarity ties an important behavioral cognitive process to how draconian punishment emerges, as the size of mistakes have a different meaning under similarity. We also vary the process of adoption to understand how institutional processes affect emergence, such as democracy versus giving a top-down planner the authority to choose punishment rules. We show that the process of democracy and endogenous selection of punishment may result in different policies. This could explain why severe punishment for egregious violators sometimes emerges in top-down institutions.

An overview of the structure of the remaining paper is given as follows. In Section 2 we discuss how the prior literature motivated our study and where our contribution fits in. In Section 3, we describe our model. We describe the game theoretic setting for the model, including the shape of the policy solution found in analytical results with rational, well-informed agents, and the details of fine vectors, which are the policy we use here (Section 3.1). In that section we also describe our computational agents' information, learning, bounded rationality, and preferences (Section 3.3) and the details of the political regimes of Private Provision, the Social Planner, and Competitive Direct Democracy (Section 3.4). In Section 4, we describe the agent behavior and fine vectors which emerge under these political regimes in our model, and in Section 5, we discuss those results, including comparing across regimes and to theory. In Section 6 we summarize the core findings, describe how the model could be extended in future work, and conclude.

2 Motivation and Prior Literature

Our paper participates in the literature on coordination problems and, more narrowly, the formation and utilization of institutions to solve the tragedy of the commons.

A great deal of work has been done on the role sanctioning can play in addressing the commons problem. It is well established that punishment serves not only as a powerful mitigator of unwanted behavior directly, but also as a signal of social norms and expectations of group behavior (Ostrom et al., 1992; Fehr and Gächter, 2000; Jules et al., 2020). We also know punishment of unwanted behavior can occur even in contexts without repeated interaction. This is commonly observed both in experiments and in the field (Boyd et al., 2003).

A large experimental literature tells us what behavior is reasonably expected in iterated prisoner's dilemma games and CPR games where peer punishment is available. Peer punishment is found to increase cooperation (Chaudhuri, 2011; Fehr and Williams, 2018). Groups endogenously choose informal or less-than-maximal punishments and perform well (Markussen et al., 2014; Engel, 2014). Many studies indicate that people are inclined to punish free-riding partners, which indicates that peer punishment can lead to a welfare enhancing cooperation (Gächter et al., 2008; Raihani et al., 2012). Many people are "conditional cooperators" in these games and use punishment to sustain cooperation in groups (Chaudhuri, 2011).

Some experimental work focuses on the degree of punishment. De Geest and Kingsley (2021) find that human subjects target egregious cheating by applying greater punishment in such cases, which suggests that small errors or 'mistakes' are tolerated, similar to the principle of graduated sanctions. Andreoni and Gee (2012) and Liu et al. (2020) find that centralized punishment of the lowest contributors is also effective as a punishment rule when the size of free-riding is incorporated into the punishment. This is important in social dilemmas where information is incomplete and small 'mistakes' may occur on accident rather than through intentional free riding.

Other experimental work focuses on the institutional design of punishment. De Geest and Kingsley (2021), mentioned above, is a study of coordination around peer punishment. People appear to prefer coordinated punishment institutions (Molleman et al., 2019), meaning punishing a free-rider when others are also punishing that same free-rider, indicating a preference for punishment rules that are popular among resource users. Groups will sort into prosocial groups when given the opportunity and adopt institutions (Fehr and Williams, 2018). Nicklisch et al. (2016) finds that top-down sanctioning may be preferred when observability is low in a public goods game.

This recent interest in institutional design dovetails with Elinor Ostrom's seminal work on CPRs (Ostrom, 1990). Across communities that varied greatly in size, geography, culture, and resources, she identified a number of features of governance structure that were frequently held in common. These features are referred to as "Ostrom's design principles." These principles of long-enduring CPRs appeared even in communities in relative isolation from one another, suggesting that they may have emerged independently.

Graduated sanctions are one of Ostrom's design principles. In the hundreds of case studies Ostrom analyzed, she observed graduated sanctions frequently and found them to be important to successful sustained management of CPRs (Ostrom, 1990). Many later works have bolstered the evidence for this, including Bardhan (1993), Ostrom (1993), Ghate and Nagendra (2005), Rubinos (2017), and van Klingeren and Buskens (2024). The design principles appear to succeed best when there is congruence between local conditions and rules and proportionality between investment and extraction (Baggio et al., 2016). Iwasa and Lee (2013) show in a theoretical model that graduated sanctions work best when the probability of erroneous reporting of players' actions is low and there is heterogeneity in the sensitivity to differences in payoffs. van Klingeren and Buskens (2024) find in experimental work that graduated sanctions are more effective than strict sanctions in the long term, and that there are specific conditions when graduated sanctions are effective. There is also historical evidence that graduated sanctions are not always necessary when other institutions

are used (De Moor and Tukker, 2015; De Moor et al., 2021).

The conditions under which particular sanctioning types emerge remains unclear. We know, for example, sanctioning behavior can be driven by inequality and reciprocal motives (Visser and Burns, 2015). It can also be driven by the type of resource being governed. Additionally, the policy that emerges must in part be a function of the deliberative process by which policies themselves are formed (De Geest and Miller, 2023). Ostrom (2000), Janssen and Ostrom (2006), and Wilson et al. (2013) all illuminate the important role that agent-based models could play in capturing such a phenomenon, which serves as the groundwork for this endeavor.[1]

Experiments and case studies have identified many important features of peer punishment. Yet studies using these methods are limited in their ability to investigate precisely which institutions emerge and how they do so. Such studies generally cannot explore the evolution of selected policies over many generations under a variety of conditions. But a game-theoretic model that incorporated all the institutional nuances identified by the literature on peer punishment would be extremely unlikely to have a closed-form solution. For these reasons, we pursue a computational approach of agent-based modeling here. Ostrom's own work suggests that evolutionary agent-based models may be well-suited to capture some of this adaptive process (Ostrom, 2000; Janssen and Ostrom, 2006; Wilson et al., 2013). Agent-based modeling seems especially appropriate because Ostrom's design principles were produced by deliberative processes among agents who devised their own policy solutions. Our work is also similar in spirit to Bowles and Choi (2013), which studies how the institution of private property emerges using computation methods, and Oliveira (2023), which describes the rise of social inequality in a pie-sharing game. If design principles like graduated sanctions tend to improve management of CPRs, then it should be possible to see the emergence of these principles in an appropriately-defined, adaptive, and evolutionary

[1]In a study of South Korean fishermen, it was found that graduated sanctions were needed to successfully manage mobile marine species, but not needed for successful management of non-mobile marine species (Shin et al., 2020).

agent-based model in which agents interact with a CPR.

Much has been done on modeling of coordination behavior in commons games with computational methods. For example, Waring et al. (2017) propose a multi-level selection model, similar to Traulsen and Nowak (2006), of resource harvesting with the aim of understanding when sustainable practices emerge as the dominant paradigm in their context. A number of papers explore evolutionary games of coordination for sustainability, including Sethi and Somanathan (1996), Tavoni A and S (2012), and Schlüter Maja and Simon (2016). De Geest and Miller (2023) explore how social-choice mechanisms affect the policy that emerges from agents playing a public goods game. Finally, and perhaps most similar in aim to ours, Couto et al. (2020) propose an evolutionary game aiming to understand why graduated sanctions are so effective, though they investigate policy graduation in the number of bad actors instead of punishment for the size of the violation as we do in this study.

Our model is distinguishable from the prior literature we have discussed here in a few important ways. First, instead of *population-level* selection methods with random mutations to strategies (e.g. the genetic algorithm, evolutionary games with replicator dynamics, etc.), we utilize *agent-level* learning through reinforcement. While population learning methods are certainly appropriate and powerful for many applications, we felt such methods may serve as insufficient for our particular question. Our initial motivation was that graduated sanctioning could emerge as helpful for learning agents in particular, since the smaller punishments could serve as "little lessons" for the agents to learn from. Given this, we felt the agents needed the ability to respond to and have their behavior guided by their own individual experiences and that a well-performing emergent policy needed to take into account that each agent has their own learning path which unfolds in tandem with the other agents. Therefore, in this context, we believe using a population-level learning method is inappropriate. This distinction is reinforced in (Ostrom, 2014) which makes the case that the evolution of rules may follow different selection processes than biological selection. Additionally, some population-level learning methods like the genetic algorithm, which rely on killing low per-

forming strategies and/or replicating high performing ones, often have limited capacity to memorize. Cycles in behavior in such models can often be thought of as the population learning the same lessons repeatedly, after their influence on the current state of the population dwindles. While this can be a desirable feature in some contexts, particularly when investigating the rise and fall of communities, we are less interested in what policy solutions can be forgotten and more interested in if the solutions could emerge at all in the first place. Ostrom (2014) also posits that rule selection or changes to rules may be viewed in some ways as a type of "policy experiment." Taking inspiration from these words, we decided to model this process of rule selection and learning based on "policy experiments" to find the better performing rules explicitly. This process applies not only to the agents' common resource use choices in the model, but also the social planner's policy choice (the Social Planner Regime) and citizen/agents' voting in representative democracy (the Competitive Direct Democracy Regime). Further, we allow our policy maker(s) and agents to have full access and flexibility to utilize all combinations of their policy and/or action spaces. We believe modeling both policy choice and agent behavior as a relatively unconstrained and intentional process of experimentation could be an important component to understanding emergent behavior in such systems.

3 Model

This is an agent-based computational model of a common resource with boundedly rational computational agents with limited information.

Agent-based models can be thought of as a form of computational game theory, in which game-theoretic agents with utility functions, available actions, information, and, here, the ability to learn computationally, are placed in a game-theory-style game. In order to see what the model predicts, instead of solving for a kind of equilibrium, one runs the computational model and learns outcomes statistically via the synthetic data generated by the model. Those

synthetic data can be evaluated econometrically.

In this simulation, N agents play a repeated common resource problem. In each period, each agent i first selects a harvest level h_i for a set of possible harvest levels $\mathcal{H}$, and second receives a corresponding payoff from the resource; an individual benefit, a common cost, and a fine (which may be zero). These agents can learn from their experiences to modify their choices.

The common resource problem specifically is the Harvest Game, a standard negative externality "tragedy of the commons" game evaluated by, among others, Ostrom (e.g., Ostrom et al., 1994) The game, its analytical solution for fully informed, rational agents, and the definition of the policy, fine vectors, is found in Subsection 3.1. The conditions under which fine vectors satisfy Ostrom's "graduated sanctions" are described in Subsection 3.2. The behavioral/computational components of the individual and social planning agents, such as learning and preferences, are described Subsection 3.3. The three political regimes–private provision, the social planner, and competitive direct democracy–are described in Subsection 3.4.

Creating a version of a model which aligns with an existing model is called "docking" in the agent-based modeling literature (Axtell et al., 1996). This particular version of docking creates an agent-based version of an analytical model. This is often called agentization Guerrero and Axtell (2011).

The code for this model is publicly available at:
`https://github.com/chriszosh1/EvolvingSustainableInstitutions`.

3.1 The Harvest Game and Fine Vectors

The core of our model is The Harvest Game, an N-player common resource that provides positive private benefit and negative externalities. It is a canonical N-player investment game with negative externalities, such as found in Ostrom et al. (1994). Each time the Harvest Game is played, each agent i must decide their *harvest level* $h_i \in \mathcal{H}$, where $\mathcal{H}$ is the set

of possible harvest levels. $\mathcal{H}$ is either a continuum $[0, H]$, for the analytical analysis which appears in this Subsection, or a discrete set of weakly positive integer values $\{0, 1, \ldots, H\}$, for the computational model.

Furthermore, we define $\vec{h}$ as a *profile* of harvest levels, i.e., a set of harvest levels, one for each agent:

$$\vec{h} = (h_1, h_2, \ldots, h_N)$$

Each unit harvested by agent i provides 1 unit of benefit to player i, but contributes to a cost which each agent must pay. The cost is driven by average harvest:

$$C(\overline{h}) = \alpha\overline{h} - \beta\overline{h}^2 \tag{1}$$

$$\text{where} \qquad \overline{h} = \frac{1}{N}\sum_{h \in \vec{h}} h \tag{2}$$

α and β are both positive parameters between 0 and 1. They denote the severity of the negative externalities. In particular, the higher α is, the more severe the negative externality, and the higher is β, the more the externality accelerates are more of the resource is harvested.

In addition to the benefit and cost, there is also a set of fines f, which is a utility cost extracted from the agent i for a choice of harvest h. f is a function $f : \mathcal{H} \to [0, M]$ for some $M > 0$, where $f(h)$ is the fine associated with selecting harvest level h. When H is discrete, f can be represented by a vector $(f_0, f_1, \ldots, f_H)$, where each $f_h \in [0, M]$.

For a continuum of possible h, f is represented as a function $f(h)$, and for finitely many possible h (as occurs in the computational model), f is represented by a vector, one entry for each possible value of h. Fine vectors under private provision with no government intervention are represented by the fine function $f(h) = 0$ for all h. Importantly, the fine function/vector is represented in this way so as to be completely flexible. It allows for arbitrary policy shapes to emerge, graduated, draconian, or otherwise. (More on this in Subsection 3.2).

Combined, the payoff to an agent i as a function of their harvest choice for a given fine vector f is

$$\pi_i(h_i, \overline{h}) = h_i - C(\overline{h}) - f(h_i) \tag{3}$$

Consider a setting with no fines (so $f(h) = 0 \forall h$). In this setting, the competitive equilibrium (i.e. Nash equilibrium) level of harvest is

$$h_{CE} = \frac{N - \alpha}{2\beta} \tag{4}$$

If one maximizes social welfare (that is, the sum of individual welfare pi_i), one can find the socially optimal harvest level:

$$h_{SO} = \frac{1 - \alpha}{2\beta} \tag{5}$$

As is expected with negative externalities, $h_{SO} > h_{CE}$.[2] It is the convention to refer to this socially optimal level as the one selected by a (benevolent) social planner.

A standard solution to negative externalities is to levy fines against behavior which deviates from the socially optimal level, so-called Pigouvian taxes (Pigou, 1929). In this setting, this places fines on all levels $h \in (2, 5]$. There are many such fine vectors which solve this problem. The minimal fine vector–that is, the smallest fines at each value of h required to achieve this result–is:

$$f^\star(h) = \begin{cases} 0 & \text{if } h = h_{SO} \\ \max(0, A + Bh + Ch^2) & \text{otherwise} \end{cases} \tag{6}$$

[2]Note that in this model, the more agents in the model, the higher is Nash Equilibrium harvest choice. This is due to smaller contribution of each agent to the average. Population growth is not a focus of this model, however, and we have kept N fixed in our analysis.

where

$$A = \frac{\beta(2N-1)}{N^2} h_{SO}^2 - (1 - \frac{\alpha}{N}) h_{SO}$$

$$B = 1 - \frac{\alpha}{N} - \frac{2\beta(N-1)}{N^2} h_{SO}$$

$$C = \frac{-\beta}{N^2}$$

Fines which exceed $f^\star$ for values $h \neq h_{SO}$ also achieve the socially optimal level of harvest. That is, fines associated with any non-socially optimal choice can be raised by any amount without any change in social welfare or implied behavior by rational agents. These results are also invariant to whether these fines are redistributed back to the agents or not. This extends directly from the fact that rational agents will only ever incur the penalties associated with on equilibrium path actions.

Table 1: Parameter Values and Equilibrium Solutions

Parameter	**Value**
Number of agents (N)	4
Maximum harvest level (H)	5
Externality parameter (α)	0.8
Externality parameter (β)	0.05
Maximum allowable fine (M)	10
Solution	**Value**
Social optimum harvest (h_{SO})	2
Competitive equilibrium harvest (h_{CE})	5

Table 1 describes the parameters we selected and corresponding theoretical solutions for this model. We chose these values to provide an internal solution for the socially optimal value far from the competitive solution. An integer value aids in the agent-based implementation since the agents choose from a discretized action space. In our computational implementation, agents may select harvest levels $\{0, 1, 2, 3, 4, 5\}$. In this case, fine vectors are drawn from a fine vector space $F = \mathbb{R}_+^{H+1} = \mathbb{R}_+^6$

For these parameter values, the minimal fine vector at integer values $\{0, 1, 2, 3, 4, 5\}$ is

$$f^{\star}(h) = [0, 0, 0, 0.746875, 1.4875, 2.221875] \tag{7}$$

This means a fine vector which extracts, for example, fines of size [10, 10, 0, 10, 10, 10] solves the problem just as well as the policy function [0, 0, 0, 1, 2, 3]. What we can conclude from this model is that, from the perspective of rational agents, the shape of the policy itself is fairly inconsequential.

That said, the game can be modified slightly so that the minimal fine vector is optimal. For example, if these rational agents faced a small chance of error, i.e. "trembles" (Selten, 1975), then the fine vector $f^{\star}$ above is the uniquely optimal fine vector. Simply put, if there is a small but positive chance that the fines will be paid, agents now prefer they are as small as possible. The desirability of Trembling Hand is similar to the desirability of Subgame Perfect Nash Equilibrium: we wish agents not only to have rational demands *on* the equilibrium behavior, but also *off* equilibrium behavior. (See Appendix 6 for details.)

One aspect of fines is what is done with the revenue generated. Here we refer to this as fines with redistribution or without redistribution. Under fines with redistribution, the total fines collected are assumed to be costlessly redistributed to individuals. Under fines without redistribution, collected fines are assumed to be destroyed and lost to society. This might be the case if the fines are e.g. imprisonment or social sanctions, which generate no value which can be redistributed but are costly to the recipient of the fine. Mechanically, without distribution, fine revenues are excluded from social welfare, and with redistribution, they are included. Note, the Nash Equilibrium above assumes fines are not redistributed.

3.2 Defining Graduated Sanctions

Under graduated sanctions, a first, small infraction might result in a small fine, but a single large infraction or repeated offenses may result in a very high penalty (e.g. banishment from

the group and the resource). Here, we consider any harvest level above the socially optimal level as an infraction, and we would describe a policy function as *graduated* if it is upward sloping and convex (so both the overall and *marginal* penalty is increasing). Furthermore, we would expect the size of the penalty at a harvest level just above socially optimal level to be small relative to the minimum size that theory predicts a penalty can take while still correcting behavior to the optimal.

Graduated sanctions can be most easily understood in contrast with draconian sanctions. Draconian sanctions maximally (or near maximally) punish players for choosing non-socially optimal harvest levels.

Based on the theoretical solution provided above, when fines are not redistributed, the planner should utilize the solution found in Equation 7. This is not graduated, as it is not convex in the region where harvest levels are above socially optimal (i.e. where $h_i > 2$). In fact, it is slightly concave.

When the planner utilizes redistribution of fines in the theoretical model, however, the social planner views the choice between graduated sanctions and draconian sanctions inconsequential, so long as the fine vector satisfies the criteria outlined in 6. This means that while graduated sanctions could solve the planner's problem, so could many other types of policies including the draconian $f(h_i) = [0, 0, 0, 10\ 10, 10]$ and the more peculiar $f(h_i) = [10, 10, 0, 7, 5, 3]$. This is because the fines result in no net loss of social welfare but provide sufficient deterrence.

Much of Ostrom's work demonstrates that these successful policies are not only graduated in the size of infractions, but also in the number of infractions (ie. different penalties for 'repeat offenders'). In our model, we only explore graduation of policy in terms of the size of infraction. Extension of the model to allow for policies with dynamic punishments is certainly an avenue for future exploration.

3.3 Information, Learning, Bounded Rationality, and Preferences

The agents in this computational model are boundedly rational and learn through reinforcement with similarity. Reinforcement learning has been well-established in economics. Originally introduced by Erev and Roth (1998), it has been shown to be effective in learning in coordination games (Sarin and Vahid, 2001), market entry games (Erev et al., 2010), ultimatum and best-shot games (Erev and Roth, 2014), and binary choice tasks (Nevo and Erev, 2012). Selten and Chmura (2008) tested a set of twelve games and found reinforcement learning was somewhat effective, although it was outperformed by some more complicated models, such as Experience-weighted Attraction (Camerer and Ho, 1999) which includes reinforcement learning as a special case. Some of the empirical evidence comes from agent-based models brought to data. For example, Kirman and Vriend (2001) use a reinforcement-learning agent-based model to describe empirically observed patterns of buyer/seller matches in a Marseilles fish market, and Duffy (2001) shows reinforcement learning agent-based models can explain currency speculation and asset market trading experiments. Reinforcement learning has also been used extensively in agent-based simulation as we do here, for example in the simulation of the El-Farol problem, a famous problem in complex systems, predicting attendance at a bar in downtown Sante Fe (Franke, 2003), and a pie-sharing game used as a metaphor for social inequality (Oliveira, 2023), as well as agent-based models of financial markets (LeBaron, 2001; Farmer and Joshi, 2002) and labor markets (Neugart, 2008).

The reinforcement process used in this paper is:[3]

1. Agents are initialized with a high attraction S to each action.

2. Agents store the average performance of each action as their attraction to each action. When an action is attempted for the first time, its performance replaces the initial attraction score assigned. When agents utilize similarity, nearby actions also have their performance updated to a lesser degree.

[3]For more details, see Appendix 6.

3. Agents have a probability p_t to explore and probability $1-p_t$ to exploit, where p_t starts large and shrinks to zero.

4. When agents explore, they choose an action with a probability proportional to its attraction.

5. When agents exploit, they choose the action with the highest attraction.

The performance of actions are determined in one of two ways. Selfish agents use own utility as their measure of performance, given in Equation 3 above. Altruistic agents use the sum of all agent's utility payoffs as their performance metric.

Unlike in traditional reinforcement learning where agents would use cumulative utility–action scores which are additively updated by experiences as seen in Erev and Roth (1998)–our agents use the additional experiences to update their estimates of each action's average performance. This is more akin to an expected utility formulation, and thus allows for more direct comparison to traditional utility formulations which agents leverage in the theoretical models.

Agents also utilize similarity as a baseline. Simply put, agents believe similar actions have similar consequences. Mechanically, when an agent chooses an action and receives feedback on the performance of that action, the agent also updates their estimates of the average performance of nearby actions to a lesser degree. This allows for agents to extrapolate from some actions to others. That is, when the agent gets, for example, a low utility payoff for choosing 4, an agent using similarity would also lower the expected payoff of 3 and 5 (to a lesser degree than 4), but not, for example, lower the expected payoff associated with 1. On the other hand, if agents were not to use similarity, then each action would be independent in the mind of the agent, and learning that 4 had a low payoff would tell the agent nothing about the payoff of 3 or 5.

The role of similarity in experiential learning is important both from an *a priori* philosophical understanding of learning, as well as an evidence-based, psychological understanding

of learning. From a philosophical perspective, Gilboa and Schmeidler (1995) quote Hume (1777) to motivate their use of similarity in case-based decision theory: *"From causes which appear similar we expect similar effects. This is the sum of all our experimental conclusions."* Similarity, in this sense, is known as the the 'Law of effect' or generalization (Guttman and Kalish, 1956; Skinner, 1957; Brown, 1958). It has been rigorously tested and established empirically, most famously by Shepard (1987), where he synthesizes data from extrapolation to establish a law of generalization, which is "invariant across perceptual dimensions, modalities, individuals, and species." Similarity is a cornerstone of case-based decision theory, which has not only a theoretical literature (e.g., Gilboa and Schmeidler, 1997; Tsatsoulis et al., 1997; Dubois et al., 1999; Gilboa and Schmeidler, 2001; Gilboa et al., 2002; Eichberger and Guerdjikova, 2020), but also a variety of empirical applications in which it was shown to be effective at describing human choice (e.g., Ossadnik et al., 2013; Pape and Kurtz, 2013; Guilfoos and Pape, 2016; Kinjo and Sugawara, 2016; Radoc, 2018; Thomas and Guilfoos, 2023).

Similarity is not just an appealing attribute in modeling agents' learning; it also aids computational tractability. If the action space is arbitrarily large, similarity may be helpful for learning to occur in a reasonable amount of time. For example, in a model of learning agents facing a continuous action space, treating each action as unrelated is infeasible.

This decision-making process strikes a balance between agent sophistication and simplicity. Our agents are sophisticated enough to engage with a highly unconstrained policy function while simple enough to facilitate interpretability of decision making and direct comparison to theory.

In this model, agents seek to maximize their own contemporaneous utility, which is derived from private consumption and the good harvested from the common resource, minus fines. At times we consider "altrustic" agents, who seek, instead, to maximize social welfare (the sum of individual utility). Altruistic agents are used to benchmark the learnability of socially optimal outcomes. When relevant, standard agents who seek to maximize their own

welfare are called "selfish" to contrast from altruistic agents.

3.4 Regimes

There are three political regimes under which the individuals in the model select their harvest levels. These regimes reflect the process by which policy is determined. In this model, the fine vector is the only policy choice, but it could be extended to consider other types of policy with these same regimes. We define these regimes for the agent-based computational model, so we refer to the agents learning.

The Private Provision Regime (Subsection 3.4.1) represents the competitive state without coordination; anarchy. It has a fine vector where all entries are fixed to be zero and cannot be changed. The Social Planner Regime (Subsection 3.4.2) has a single, benevolent social planner who selects the fine vector after repeated trials. The Competitive Direct Democracy Regime (Subsection 3.4.3) has two political actors–political parties, perhaps–which each simultaneously offer alternative fine vectors to the voting public, who update their opinions on the relative merits of these fine vector outcomes based on limited trial runs.

Other regimes can be implemented in this framework. We discuss this idea briefly, with examples, in our conclusion and future work (Section 6).

3.4.1 The Private Provision Regime

The Private Provision Regime is simple: the fine vector is simply zero for all harvest values. This represents the state of a free market with no government intervention and no explicit coordination among agents. The agents cannot, for example, engage in private agreements of any kind. Nothing rules out implicit coordination, however, as one might see in a Nash equilibrium. Since the agents learn payoffs via trial and error (see Subsection 3.3 for more details), it is not a foregone conclusion whether they converge to the Nash Equilibrium for fully informed agents as described in Subsection 3.1.

3.4.2 The Social Planner Regime

In the Social Planner Regime, we introduce a benevolent social planner who cares equally about the welfare of all agents and selects policies. This model modifies and extends the Private Provision Regime above. Other than variation in policies, the model is unmodified. Importantly, the citizen agents in the model remain the same: they are the boundedly rational learning agents described in Subsection 3.3. They learn under a fixed policy for a set number of periods, after which the Social Planner measures cumulative utility. The Social Planner can then try a new policy with a reset set of agents; agents who have no recollection of previous policy experiments.

As described in Subsection 3.1, the policies are fine vectors f drawn from $F = \mathbb{R}^6_+$.

Social welfare is given by

$$\Psi(\vec{h}) = \Sigma_{i=1}^{N} \Sigma_{t=0}^{T} \pi_i(h_i, f(h_i)) \tag{8}$$

$$\text{where} \qquad \pi_i(h_i, \overline{h}) = h_i - C(\overline{h}) - f(h_i) \tag{9}$$

$$\text{and} \qquad C(\overline{h}) = \alpha\overline{h} - \beta\overline{h}^2 \tag{10}$$

$$\text{and} \qquad \overline{h} = \frac{1}{N} \sum_{h \in \vec{h}} h \tag{11}$$

Lines 9, 10, and 11 are defined in Subsection 3.1 Equations 3, 1, and 2.

The social planner computationally explores the fitness landscape of the fine vector space via a combination of hill climbing and simulated backwards induction, which we aim to summarize as follows:[4]

The social planner starts with a fine vector f from uniformly from the fine vector space $F = \mathbb{R}^6_+$ (including decimals up to hundredths). Each round of the simulation, the social planner compares their stored 'best so far' fine vector to close alternative policies in F by running a simulation of the world under each candidate fine vector. The planner then observes the social welfare generated under each fine vector and keeps the best from that

[4]For more details, see Appendix 6.

set of options.[5] This comparison to neighboring policies is repeated once for each depth of the search we allow the social planner. Through this iterative process, the social planner explores the fine vector space with intention to find a high-performing fine vector. Since this process is path-dependent, we also allow the social planner to repeat this whole process a number of times (in our case, 25 times) from different starting points (i.e. starting with different randomly chosen initial policies). At the end of this process the best fine vector which they have found so far we denote $f^{\star}$. To ensure this exploration process was not ended prematurely, we look at the marginal improvements in social welfare over search-time of the fine vector space to look for convergence.

This process is analogous to backwards induction. We can think of this process in terms of a two-stage game in which the social planner must choose a fine vector first. After the fine-choice stage, the agents then decide, in the next stage, how to harvest in response to the fine vector. It can be thought of as the social planner running repeated "internal" simulations of agents' responses to the policies, until finally deciding to implement the policy which appears to perform best. Once the policy is decided upon, the policy is realized and agents must now decide how to respond to it. This approach allows the social planner to engage with the complicated optimization problem without encoding policy preferences.

Much like how Nash equilibria can be thought of as fixed points resulting from some dynamic learning process, the policy solution the social planner converges on at the end of their search can also be thought of the same way. The final solutions can be directly compared to the equilibrium solutions found in the Nash Equilibrium analytical results in Subsection 3.1 for fully informed agents.

3.4.3 The Competitive Direct Democracy Regime

In the Competitive Direct Democracy Regime, we investigate the effect that social choice mechanisms have on the shape and efficiency of emergent policy when we replace the social

[5]Since any digit with two decimal places from [0,10] can take a position in the fine vector, there are 1000 possible values for each of the six harvest levels, yielding 10^{18} possible combinations.

planner with an implementation of a two-party representative democracy with faithful representatives. The parties put forth policy proposals which, if elected, they implement. They adjust those policy proposals over time in an attempt to win over their opponent party. We call this Competitive Direct Democracy because it can also be thought of as the citizens voting directly on the policy, but that there is only a small menu (two) policies available, and the policies are managed by parties seeking victory. This model extends and modifies the Private Provision Regime and Social Planner Regime above. The setting is identical, and, importantly, the citizen agents in the model remain the same: they are the boundedly rational learning agents described in Subsection 3.3.

In this representation of democracy, all agents participate in shaping policy by voting over fine vectors. Each election process can be summarized as follows:[6]

1. To begin, each party proposes a fine vector drawn randomly.

2. Agents forecast the effect each fine vector will have on their own expected utility by running a small number of simulations using the model.

3. Agents vote for the policy which they forecast will yield themselves the highest expected utility, with majority rule deciding the winning policy.

4. The losing party amends their position in policy space informed by change in vote-share they received and then chooses a new policy nearby it in a manner akin to hill-climbing. The winning party maintains their current platform.

5. Moving forward, each round, agents forecast the effect the new fine vector will have on their own expected utility and compare it to the incumbent policy by running a small number of simulations using the model. We return to Step 3 and repeat.

In this model of democracy, parties are free to adopt different platforms. The representatives do not value the welfare of the agents directly; instead they only care about getting elected (and, to a lesser extent, their vote share).

[6]For more details, see Appendix 6.

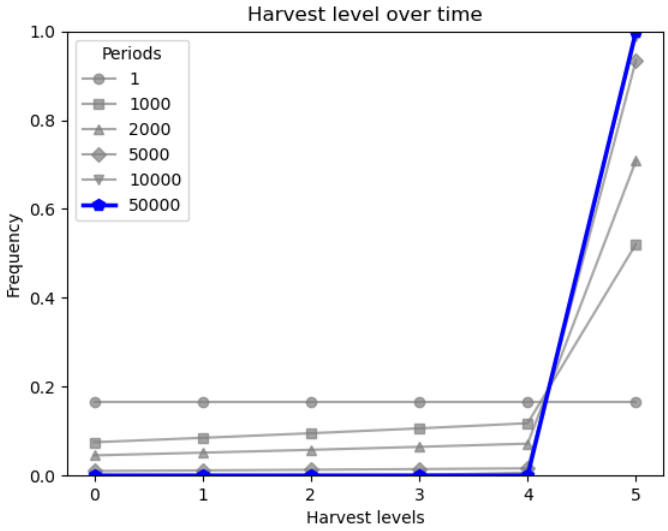
4 Results

In this section we present the model results for all regimes and variations. In Subsection 4.1, we present the private provision regime results. In Subsection 4.2, we present the results under a social planner, both without (Subsection 4.2) and with (Subsection 4.2) redistribution. Finally we present the results under Competitive Direct Democracy in Subsection 4.3. In Section 5, we summarize the results in Table 2 and compare, contrast, and discuss the results in greater detail.

4.1 Private Provision Regime Results

Figure 1 depicts the average probability of selecting each of the six available harvest levels across 25 runs at different periods of the Harvest Game. We see agents fairly quickly converge to the maximum harvest level of 5, demonstrating the tragedy of the commons. This behavior is consistent with theory. Recall the Nash equilibrium level $h_{CE} = 5$. Sice these agents act identically to what theory predicts in the long run, so agent behavior in the Private Provision Regime docks fairly closely to what theory predicts.

Figure 1: Harvest Levels in the Private Provision Regime



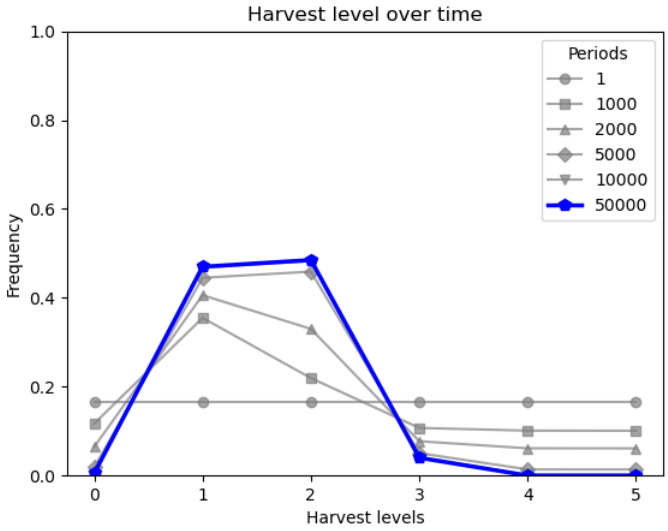
Agents learn to play the individually optimal choice in absence of policy.

Agents achieve a level of social welfare similar to theoretical results, with $\overline{\Psi}(f(.), X) \approx -0.23$ (The prediction from theory is -0.25). We can use Equation 12 to see how much social welfare is recovered (if any) when selfish learning agents are left to their own devices. We find on average, only about 4.3% of the social welfare gap is recovered.

Altruistic agents. Now consider altruistic agents, who individually seek to maximize social welfare instead of individual contemporaneous utility (see Section 3.3). Recall the socially optimal choice is $h_{SO} = 2$.

We find a population of all altruistic agents chooses a harvest level of either 1 or 2 frequently. The plot below in Figure 2 illustrates the distribution of the agents' attraction to actions as it changes over time.

Figure 2: Harvest Levels in the Private Provision Regime with Altruistic Agents



Altruistic agents learning to play pro-socially.

We note (and can observe in Figure 2) agents' attraction to 1 is of non-trivial size. Upon further investigation, we can identify the attraction to 1 comes from early periods of play where choosing 1 is a good strategy to offset other volatile agents who at times may harvest at levels higher than socially optimal. If you look at Figure 2, you can see this dynamic occurring - agents have accumulated much of their attraction to 1 by period 2000, where as agent attraction to 2 still grows to a fairly large degree from periods 2000 to 5000. Further, you can see the attraction to 1 becomes less predominant relative to 2 as attractions to actions greater than 2 decreases. This is consistent with behavior we expect from trembling hand agents, as a harvest level below 2 can be part of an optimal strategy when $\epsilon > 0$ (that is, when mistake-making occurs with some frequency). Even still, it is clear that the socially optimal choice of 2 becomes the favorite action most often, and the the difference is in the direction of pro-sociality.

We can also compare the average social welfare generated by altruistic agents in the simulation to that generated in the theoretical model. For this analysis, we interpret the

results from the following equation:

$$\overline{\Psi(X)}_{Recovered} = \frac{\overline{\Psi}(X) - \overline{\Psi}(h_{CE})}{\overline{\Psi}(h_{SO}) - \overline{\Psi}(h_{CE})} \tag{12}$$

where $\overline{\Psi}(h_{CE})$ is the average single period utility generated when agents all play $h_{CE} = 5$, $\overline{\Psi}(h_{SO})$ is the average single period utility generated when agents all play $h_{CE} = 2$, and $\overline{\Psi}(X)$ is the average single period utility generated for an agent when learning agents (who have given decision making variables X) play the game. Since our model is stochastic to some degree, we compute $\overline{\Psi}(X)$ as an average over 25 separate runs of the model.

This metric normalizes the average social welfare generated in our simulation by the gap in social welfare generated in the Harvest Game when agents act altruistically vs. selfishly. If we think about the social welfare generated when agents are selfish as reality and we consider the social welfare generated agents are altruistic as our goal, this metric will tell us how much of the social welfare gap is recovered in our simulation.

In this particular case, we are interested in seeing how closely our altruistic learning agents who use similarity get to what theory predicts social welfare should be in the absence of policy. Under the game parameters described above, player will receive a payoff of $\overline{\Psi}(h_{CE}) = -.25$, $\overline{\Psi}(h_{SO}) = .2$, and $\overline{\Psi}(X) \approx 0.187$. Thus, we find our altruistic learning agents recover about 97% of the social welfare lost when agents act selfishly rather than altruistically.

4.2 The Social Planner Regime Results

In this regime, the social planner chooses fine levels as described in Subsection 3.4.2. We consider fines without, and then with, redistribution. Without redistribution, the fines are 'pure' punishment (for example, incarceration); with redistribution, the fine revenues are put toward some socially useful purpose and are not lost. Subsection 3.1 for details.

The Social Planner Chooses Fines without Redistribution

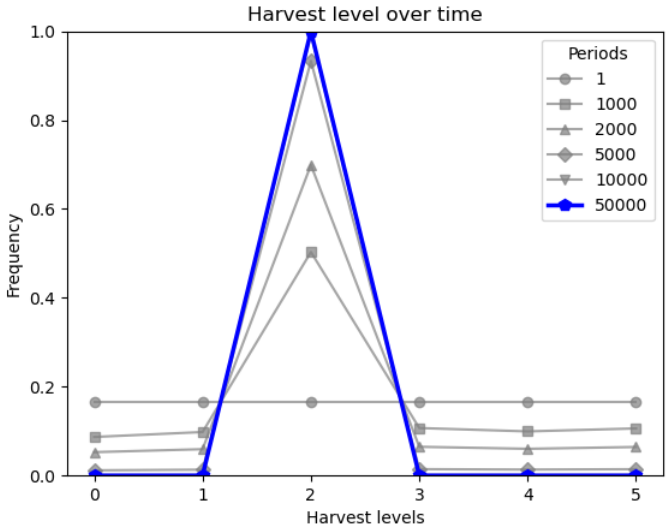
Here we consider social planning with fines without redistribution. We present the emergent policy resulting from the bounded optimization of the social planner in two contexts, one in which the agents use a particular cognitive process important to learning, similarity, and one in which they do not. (See Subsection 3.3 for details.)

When agents do not use similarity, we find the social planner converges on this fine vector:

$$f^*(.) = [0.0, 0.0, 0.0, 0.92, 2.13, 2.42]$$

This fine vector is approximately 25% above the theoretical solution given in Equation 7. First, we can see that the non-trivial portion of the fine vector (ie. the region above the socially optimal choice) is non-convex and in fact is concave, as theory predicted. Second, by looking at the plot in Figure 3 documenting agents' choices over time, we can confirm that the policy makes it incentive compatible for the selfish agents to choose the socially optimal level of harvesting $h_{SO} = 2$. This means this policy effectively implements the socially optimal level, but without graduated sanctions.

Figure 3: Harvest levels In The Social Planner Regime without Similarity



Non-similarity utilizing agents learn to play the socially optimal choice under top down policy

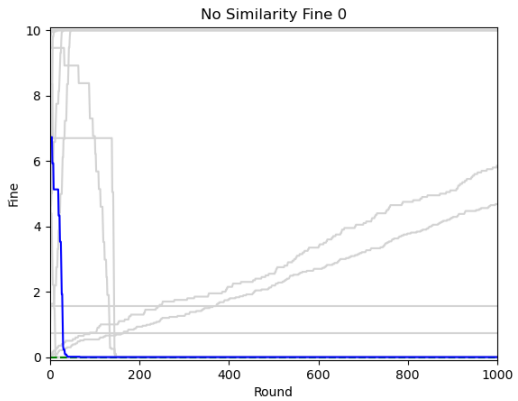
Further, we can use a generalization of Equation 12 to draw some conclusions about how well this policy solves the social planner's problem.

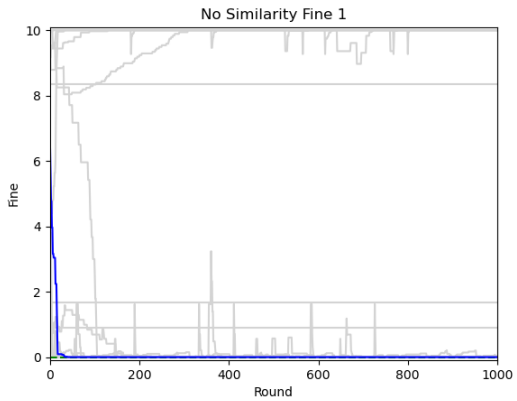
$$\overline{\Psi(f(.), X)}_{Recovered} = \frac{\overline{\Psi}(f(.), X) - \overline{\Psi}(h_{CE})}{\overline{\Psi}(h_{SO}) - \overline{\Psi}(h_{CE})} \tag{13}$$

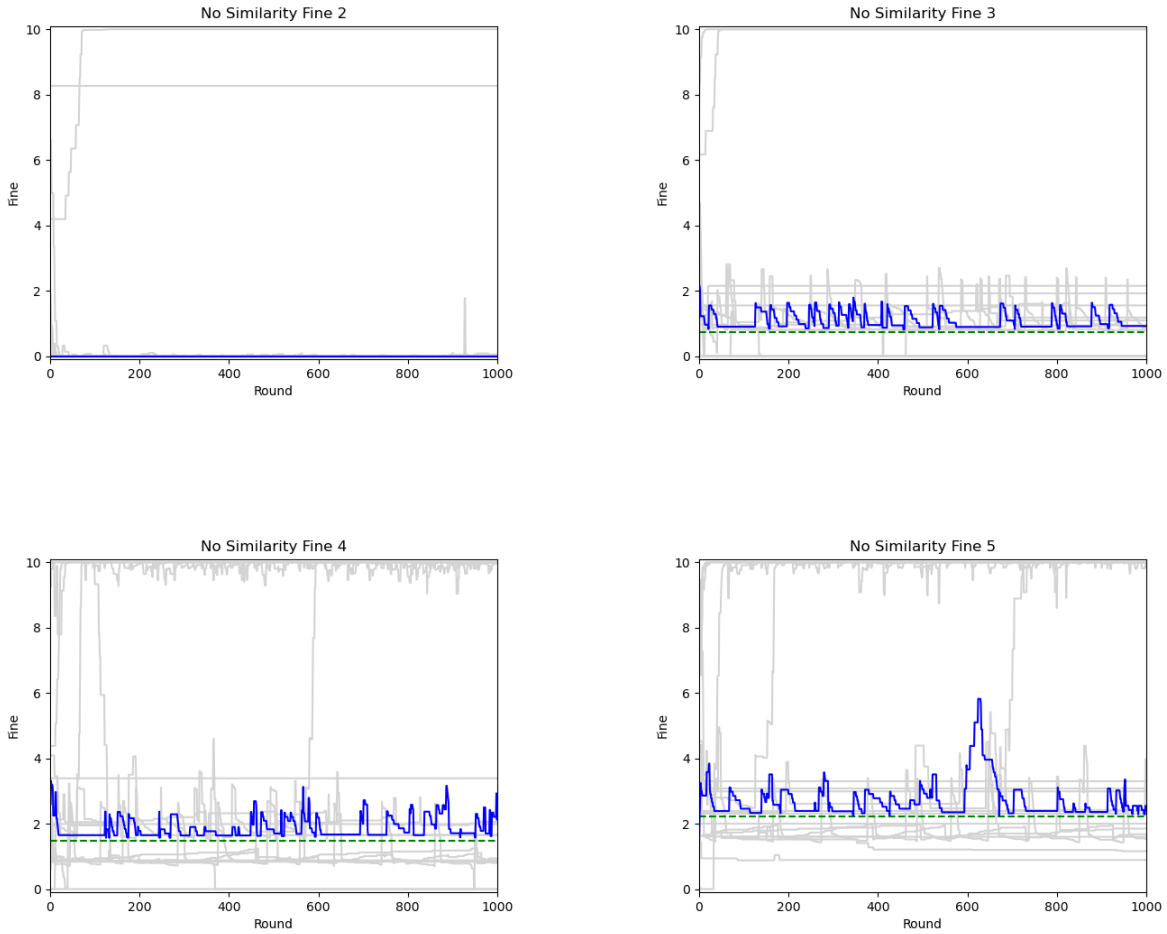
This equation will tell us how much of the social welfare lost (when agents act selfishly rather than altruistically in the absence of policy) is recovered when learning agents use X in their decision making and face policy $f(.)$. As before, $\overline{\Psi}(f(.), X)$ is computed as an average from 25 separate runs of the model. We find $\overline{\Psi}(f(.), X) \approx 0.16$ and approximately 91.3% of the social welfare is recovered by this policy.

While the policy seems to solve the problem pretty effectively, as noted above, it is non-convex. Thus the fairly effective policy solution discovered by the computational social planner does not fit our definition of graduated sanctions.

The dynamics of the social planner's policy search is presented below. Each plot corresponds to fines associated with one of the six harvest levels. Each gray line corresponds to the dynamics of fine level corresponding to one of the twenty rounds of policy searching the social planner does. The blue line corresponds to the dynamics of the policy search which resulted in the highest performing final policy.







First, we note that many of the results lock into extreme policy solutions, either employing the minimum or maximum fine allowed. This demonstrates that path dependence may limit the social planner's ability to find high performing solutions while searching locally in the policy space, validating the need to perform multiple searches. Second, we can see that while the fines associated with $h_i \in \{0, ..., 3\}$ rarely result in levels far from the theoretical lower bounds (indicated by the green dashed lines), the social planner appears to sub-optimally solve the commons problem by maximizing fines associated with $h_i \in \{4, 5\}$. Observing the dynamics which lead to the best solution (indicated by the bold, blue line), we can see that the social planner quickly reduces the fines associated with socially optimal and pro-social behavior to 0 early on and with much of the remaining time spent moving adjusting levels up and down just above the minimally corrective level of fining. We also see that the fine

levels for $h_i \in \{3, 4, 5\}$ vary regularly during policy exploration, but it always returns to approximately the same level just above the minimally corrective level.

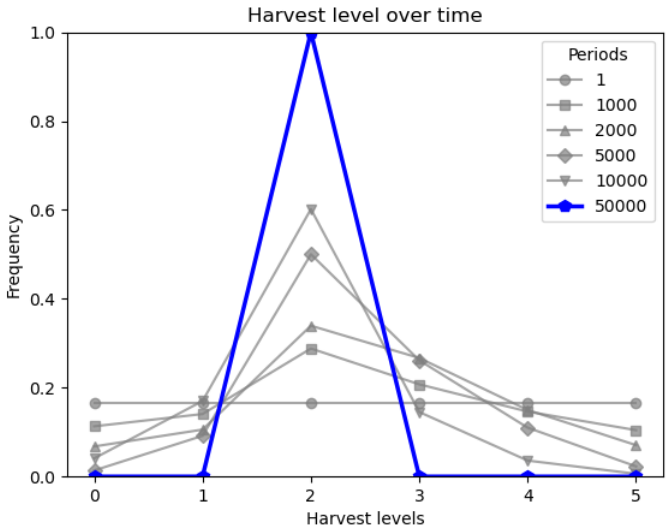
In the second case, where agents do use similarity in their decision making, the social planner finds a different-looking policy solution:

$$f^*(.) = [0.0,\ 0.0,\ 0.0,\ 0.99,\ 1.7,\ 10]$$

We can see that much of the policy looks similar to the policy solution found when agents do not utilize similarity in 4.2, with violations for choosing an h_i of 3 or 4 averaging to about 23% above the theoretical solution given in Equation 7. We find, however, that the ultimate infraction comes at a hefty fee of a maximal fine of 10.

Also as before, looking at the plots below, players quickly learn to choose a socially optimal harvest level of 2.

Figure 4: Harvest levels in the Social Planner Regime with Similarity



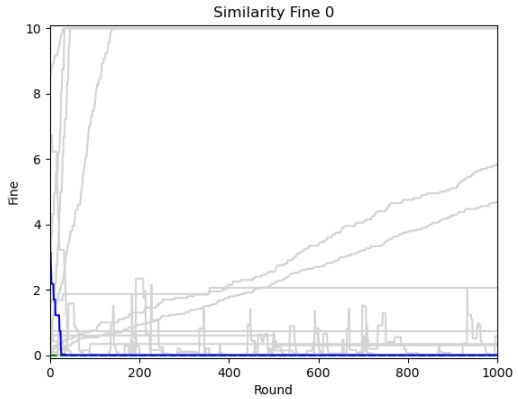
Selfish, similarity-utilizing agents learn to play the socially optimal choice under top down policy

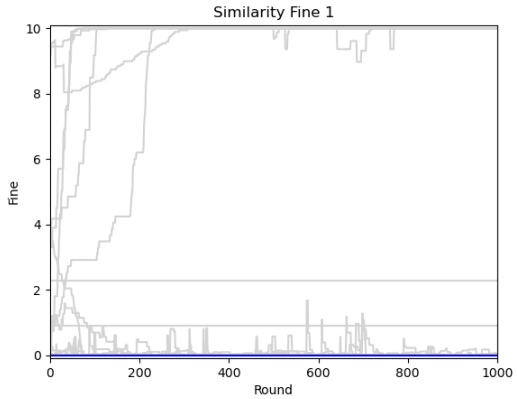
This policy nearly satisfies our definition of graduated sanctions. The best policy adopted by the social planner has a small fine for choosing the smallest violation, $h_i = 3$, only about 33% greater than what theory predicted was necessary for optimality in the trembling hand

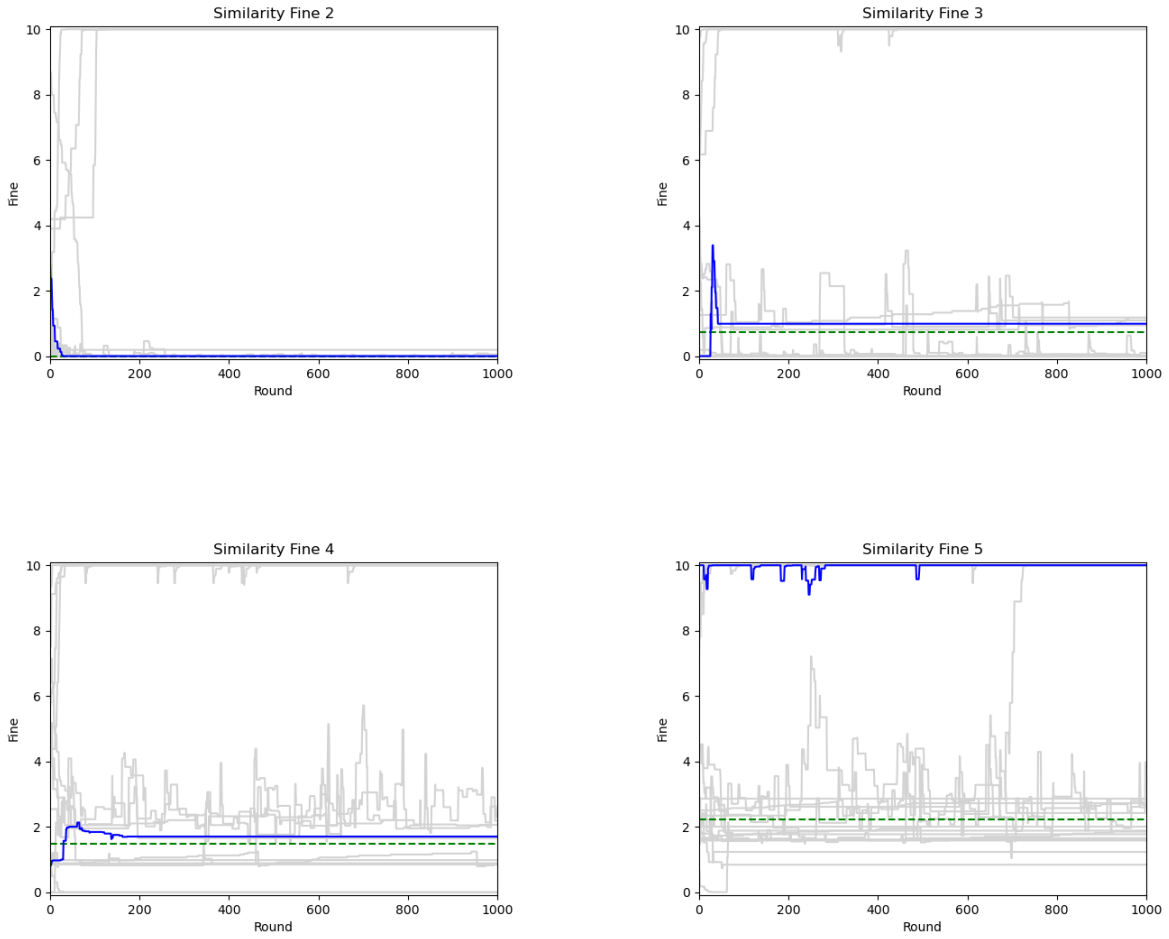
case. For the maximum violation of $h_i = 5$, however, we see a fine which is approximately 4.5 times greater than what theory predicted. This policy is not convex; the changes are $0 \rightarrow .99 \rightarrow .8 \rightarrow 8.3$, and the third change fails to be larger than the second. However, it is close to convex in that the fine for the smallest sanction is relatively small while the largest violation comes with at a hefty price.

Noting that f(.) is characterized by the policy above in 4.2 and our agents are selfish and use similarity (a characterization of X which will be used henceforth without further restatement), we find that $\overline{\Psi}(f(.), X) \approx 0.162$. Once again using Equation 13, we see that approximately 89.8% of the social welfare gap is recovered. Interestingly, this best found policy (4.2) performs almost exactly as well at recovering social welfare when agents use similarity as the previously best found solution (4.2) when agents do not use similarity. By only changing agent reasoning to utilize similarity, we see the commons problem is equally well solved but by a fairly different policy shape. Further, we see a graduated policy function emerge from top-down exploration of fine-based policies without redistribution by a social planner.

When agents leverage similarity without redistribution, we see a change in the social planner's policy dynamics.







We can see that the between search volatility seems to vary more than in the previous case. While some searches seem to not alter much, others seem to vary quite a bit. This could be due to the fact that when agents leverage similarity, changes to a fine have implications on more than one action, which in turn may alter the shape of the fitness landscape. There may be more peaks in the policy fitness landscape and perhaps regions which are relatively flat. The social planner's best found solution (in blue) appears to be fairly stable after early periods of the search. As before, in the case of the highest performing solution, the social planner quickly finds that minimizing the fines associated with $h_i \in \{0, 1, 2\}$ is best. We also do not see the cycling behavior exhibited in the previous runs during the search which produced the best performing solution.

The Social Planner Chooses Fines with Redistribution

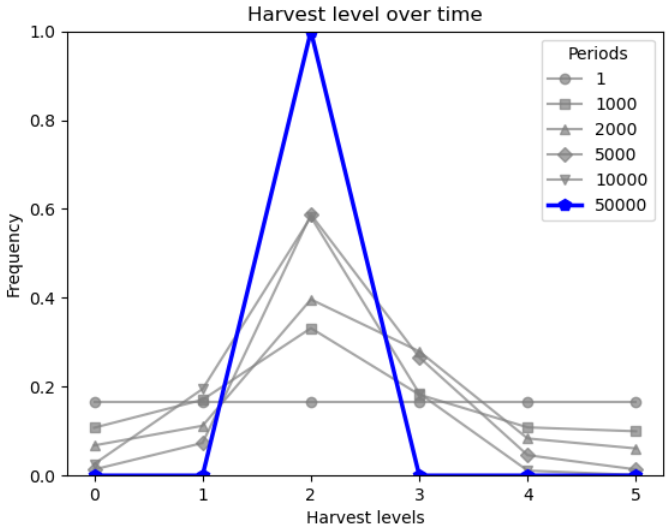
Here we consider social planning with fines with redistribution, in which the fines collected, pooled, and then redistributed, contributing directly to social welfare. Recall that, from theory, a wide array of policies could maximize social welfare equally well, as seen in Equation 6. In this context, we hypothesized that a fine vector resembling draconian sanctions would be more favorable than in the version of the model where there was no fine redistribution, as the lost social welfare from fining bad behavior reenters social welfare (one for one) through another channel. The policy the social planner discovered was:

$$f^*(.) = [0.0, 0.0, 0.0, 0.8, 10, 9.99]$$

This policy is almost draconian in that they maximally punish $h_i = 4$ and $h_i = 5$ but not $h_i = 3$. The reason is, the social planner wants players to pick the socially optimal level of investment $h_{SO} = 2$. They have no concern for how much non-compliers are punished. This leads us to believe that a draconian policy like [0, 0, 0, 10, 10, 10] might perform well, as all of the actions which agents would normally prefer over the socially optimal are now very costly to choose. Agents decide with similarity however, so choosing 3 early and receiving a huge fine may dissuade players from playing 2. Hence, we see a drop off in the intensity of fines for choosing a harvest level of 3, as it is fairly similar to the socially optimal choice of 2. Runs without similarity indicate that the similarity is pivotal in lowering the fine on $h_i = 3$ from near 10 to it is much lower level of 0.8.

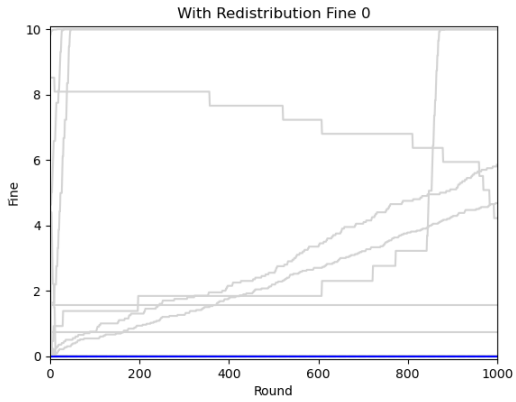
This policy choice is consistent with theory in that all of the fine levels chosen fall within the set of fine functions which will maximize social welfare as shown in Equation 6. Once again, we can see in the plot below that the best policy found by the social planner corrects behavior to the socially optimal level in the long run.

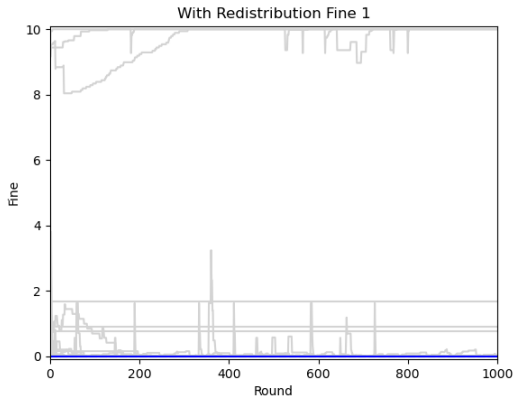
Figure 5: Harvest levels in the Social Planner Regime with Redistribution

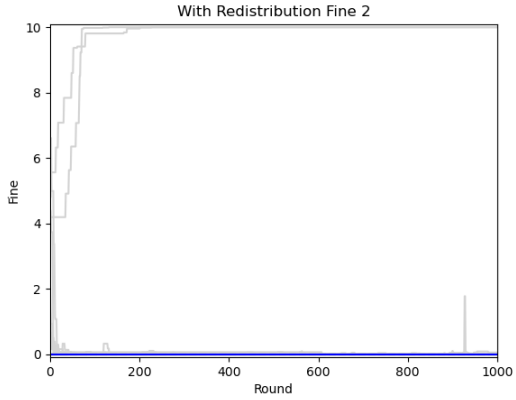


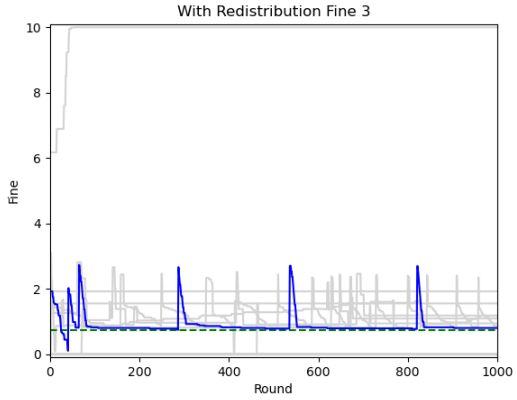
Selfish agents learn to play the socially optimal choice under top down policy when fines are redistributed.

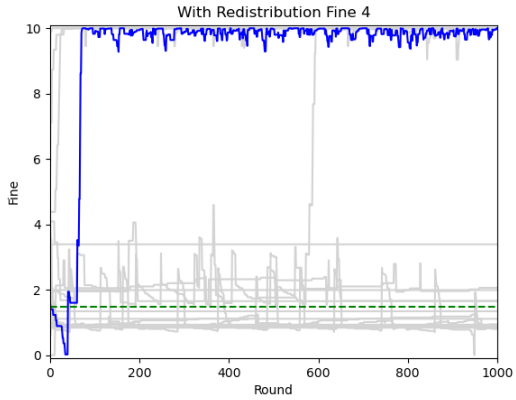
The right hand side of the plot above shows that the yellow, green, and blue lines are much closer to 0 than in the cases where fines were not redistributed, demonstrating that agents' behavior is corrected much faster with the draconian sanctions. Unlike in previous plots, by round 2000 agents almost never choose harvest levels of 4 or 5. Since fine redistribution nullifies the reduction in social welfare from punishing agents, a policy punishing agents severely has become more appealing to the social planner than it did previously. As before, we calculate $\overline{\Psi}(f(.), X)$ as the average from 25 separate runs of the model. We find $\overline{\Psi}(f(.), X)$ $\approx$ 0.195 and approximately 99.0% of social welfare is recovered by this policy when fines are redistributed. If we look at what social welfare under this policy if fines were not redistributed, we would find $\overline{\Psi}(f(.), X)$ $\approx$ -0.05 and only approximately 43.5% of social welfare would be recovered. We can see that without redistribution, social welfare would suffer quite a bit from the aggressive fining. These plots represent the dynamics of the policy search performed by the social planner in this final case.

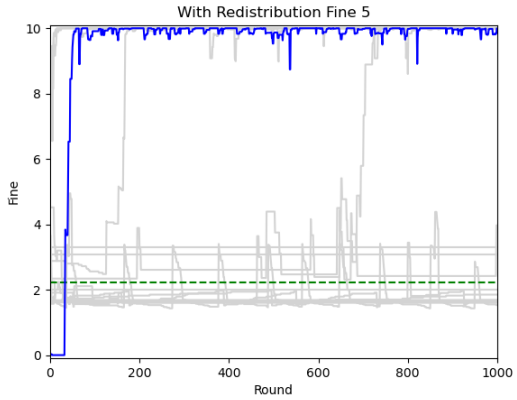












As previously observed, we can see that very few if any of the sub-optimal final policies utilize high fines on $h_i \in \{0, .., 3\}$. We also see once again that many of the sub-optimal solutions utilize fines at one of two extremes for $h_i \in \{4, 5\}$: either near the minimally

corrective level or at the maximum level allowable. Looking at the dynamics of the policy which resulted in the highest performing policy solution, we can see that while the social planner roughly converges early on near maximal fines for $h_i \in \{4, 5\}$, giving no penalty for $h_i \in \{0, 1, 2\}$, and fining near the minimally corrective level when $h_i = 3$. Once again, we can also see that the fine levels for $h_i \in \{3, 4, 5\}$ vary semi-regularly. Interestingly, the variation observed in fine associated with $h_i = 3$ seems semi-regular, as if it is cycling.

4.3 The Competitive Direct Democracy Regime Results

The dominant policy which emerges under our implementation of democracy is:

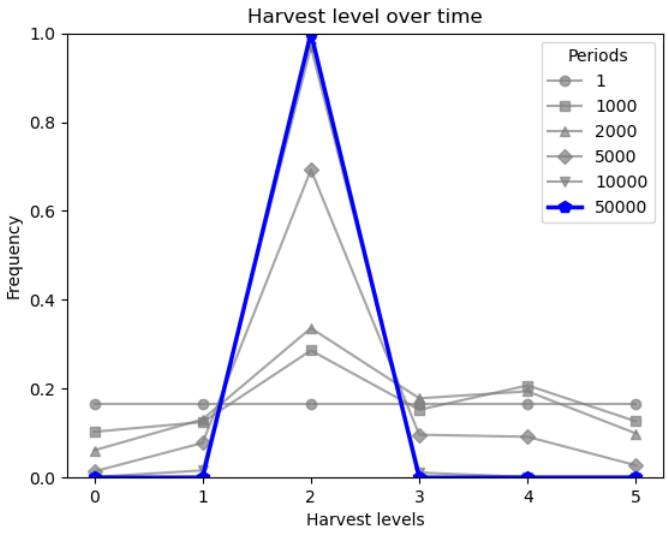
$$f^*(.) = [0.05,\ 2.7,\ 0.01,\ 2.92,\ 1.75,\ 5.53]$$

The fines in this case are not monotonically increasing in harvest levels; instead fines are high right on either side of the socially optimal level 2, suggesting a need to disincentivize small deviations from optimum, and a high fine for the individual optimal level of 5 has a high fine so that it is deterred. The fine on 4 is not as necessary, because of similarity: the fines for 3 and 5 apply heavily to 4 as well.

This carries some noticeably different features from policies which emerge from a top-down, benevolent social-planner in two important ways. Second, this policy fines players who choose less than the socially optimal level. The policy even fines players who choose the socially optimal level, though the fine is extremely small.

While it may seem surprising, this policy does fall into the large set of policy solutions that can be found in the version of the Harvest Game without trembling hands, meaning it should correct behavior (excluding that choosing $h_i = 2$ should yield a fine of 0, not 0.01, though this small fine is likely an artifact of the search process). We can see that it solves the tragedy of the commons problem by successfully encouraging agents to choose the optimal level $h_i = 2$ in Figure 6.

Figure 6: Harvest levels in the Competitive Direct Democracy Regime



Selfish agents learn to play social optimal in the face of bottom up policy.

As we did in Section 4.2, we calculate how much social welfare is recovered by implementing this policy using Equation 13. Performing this calculation, we find $\overline{\Psi}(f(.), X) \approx 0.115$, meaning that the policy recovers 81.1% of the social welfare lost when agents act selfishly rather than altruistically. Democracy fares quite well compared to the Social Planner under top-down policy (Section 4.2), where 91.3% of the social welfare loss is recovered.

5 Discussion

In this section, we summarize, compare, contrast, and discuss the results introduced in Section 4 and summarized in Table 2. Briefly, the Private Provision results (rows 1 and 2) provide a baseline in which agents largely choose to harvest $h_{CE} = 5$ and, when agents are altruistic, they recover nearly all. Altruistic private provision shows that the agents are capable of learning the socially optimal level, given the proper incentives (in this case, internalized). The first Social Planner result (row 3), with agents who use similarity and fines which are not redistributed, we see that nearly graduated sanctions emerge; for $h \geq 2$, the

Table 2: Model Results

Regime, Variations	**Fine Vector** $(f_0, f_1, \ldots, f_5)$	**Fine Vector Shape**	**% of Social Welfare Recovered**	**Section**
Private Provision	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0)	Flat	0%	4.1
Private Provision, with Altruistic Agents	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0)	Flat	97.00%	4.1
Social Planner	(0.0, 0.0, 0.0, 0.99, 1.7, 10)	Nearly Graduated Sanctions	89.83%	4.2
Social Planner, Agents without Similarity	(0.0, 0.0, 0.0, 0.92, 2.13, 2.42)	Concave for High h Near analytical	91.30%	4.2
Social Planner, Fines with Redistribution	(0.0, 0.0, 0.0, 0.8, 10, 9.99)	Draconian for High h	98.99%	4.2
Competitive Direct Democracy	(0.05, 2.7, 0.01, 2.92, 1.75, 5.53)	Large near $h_{SO} = 2$ and High h	81.07%	4.3
Minimal Analytically Optimal, for Rational, Fully Informed Agents	(0.0, 0.0, 0.0, 0.74, 1.49, 2.22)	Concave for High h	100%, 15.18%	3.1, 5.2

The modal choice under Private Provision is the individually optimal level $h_{CE} = 5$.
The modal choice in all other cases is the socially optimal level $h_{SO} = 2$.
All agents are selfish and use similarity unless otherwise noted.
All fines are without redistribution unless otherwise noted.
The fine vectors under Private Provision are fixed.
The fine vectors under Social Planner emerge from a seach process.
The fine vectors under Democracy emerge from a repeat voting process.
The minimal analytically optimal fine vector is derived from theory.
The minimal analytically optimal vector recovers all social welfare for rational, fully-informed agents, but not for computational agents. See Section 5.2.

fine function is (nearly) convex. Then there is no similarity (row 4), the fine vector is nearly concave, and is actually quite close to the minimal analytically optimal fine vector, depicted on the last row (row 7). Social planning with redistribution (row 5) yields a draconian policy for $h \geq 4$. Competitive Direct Democracy (row 6) does not yield graduated sanctions, but does preform very well, nearly reaching the level of social welfare associated with the social planning model.

5.1 Social Planning v. Democracy.

We now compare the policies which emerge from social planning versus direct democracy. The level of social welfare achieved by democracy is quite remarkable when comparing it to the top-down computational social planner who is both benevolent (caring about all agents equally) and fairly information unconstrained (having a fairly accurate forecast on how happy a policy will make each agent). Despite having purely selfish, information constrained learning agents, and despite only having two representatives to vote for, the policy found via democracy recovers about 81.1% as much social welfare as the policy recommended by the benevolent social planner.

Our hypothesis was that democracy would find a solution similar to the benevolent social planner's, both in policy shape and performance; we found it was a different shape but did not fall too short in improving social welfare. After investigating runs which allowed twice as many elections, this seems not to be the case as there was no noticeable change in the optimality of the emergent policy. The reason is as follows. In our model, representatives take the position that if they are losing, something needs to change. It is through this desire to win the election that the agents are driven to refine their policies, resembling in many ways market competition. It is also the case, however, that the winner is under no such pressure to change if their existing platform has performed very well in the past. Given this, a proposed policy that has room to improve but nonetheless dominates in the voting competition will cap out at the maximum potential of the best outside option. For example,

if party A and B have platforms in different regions of the policy space, whenever one of the parties 'peaks' (*i.e.*, finds the best solution nearby), the other also stops improving.

This is analogous to a second price auction, which provides some intuition. If A and B have two good policies, but A's policy is only slightly better than the best B can do in their region of the policy space, then A will continue to win elections with their existing policy, which means policy no longer changes. Even though A could potentially improve long-run social welfare by searching the policy space, that could risk re-election without an increase in winning frequency. So the analogy to the second price auction is, A wins as "the highest bidder" and "pays" social welfare just a hair above the social welfare generated by the best policy B can find.

In the very long run, given how we have modeled democracy, if B ever randomly approaches A in the policy space, then it is possible for democracy to find the optimal solution, as both political parties will be climbing the same hill, so to speak. This may, however, take a very long time, depending on the fitness landscape of the policy space. Additionally, this possible (though highly improbable) event of two parties advocating for similar policies may not often map back to the real world (especially given recent concerns about political polarization).

The policy discovered by democracy loses social welfare due to excessive fines, despite the fact that representatives don't directly receive these excess fines collected. Presumably this problem would worsen if they did.

We conclude that while democracy was able to solve to commons problem, it did **not** do so utilizing a graduated sanctions fine vector.

5.2 Effectiveness of the Minimal Analytically Optimal Policy

We also demonstrated that in our computational model, agent behavior matched our first set of theoretical predictions in the absence of policy in Section 4.1. We also saw in both cases that a similar level of social welfare was achieved as theory predicted. The policies

found in the Social Planner Regime (Section 4.2) and in the Competitive Direct Democracy Regime (Section 4.3) differed from the minimal analytically optimal level.

We investigate how well the minimal analytically optimal fine vector recommended by theory (see Section 3.1) performs at correcting the behavior of our selfish learning agents. We find that an agents average round payoff under this policy given by $\overline{\Psi}(f(.), X) \approx$ -0.182 and only about 15.2% of the lost social welfare is recovered. Restated, this means that the policy recommendation given above in (4.2) performs more than 6 times better than the one proposed by theory.

This may be surprising to some. Simply put, the theoretical model's agents do not make mistakes as a function of the payoffs while our learning agents do. Given higher fines will discourage learning agents from choosing certain actions as frequently, it seems fairly intuitive optimal fines may need to be higher as they play an additional role in discouraging future exploration of actions which are particularly harmful to social welfare. This increase in fines over the theoretical model's solution we noted earlier will result in some direct loss in social welfare when compared to the social welfare achieved in the theoretical model (about 8.5% to be precise), but again, is offset by the benefits from discouraging future exploration of particularly costly actions. This 8.5% social welfare loss can be thought of as a social welfare premium paid for having a population of learning agents with endogenous mistake-making / exploration.

Applying the theoretical policy solution to the computational model resulted in fairly poor levels of social welfare attainment. This exemplifies the potential sensitivity of policy solutions to cognitive simple cognitive processes. In our case, we found fairly different policy solutions for a population of agents who learn and explore actions intentionally as opposed to rational decision making with exogenously determined, uniform mistake-making.

5.3 Establishing Sufficient Conditions

Through the variations of our model that we have explored, we have started to characterize which conditions are sufficient and which features may be pivotal in determining which policy shapes produce the most social welfare in their context. In particular, we draw three primary conclusions from our simulated trials:

How agents learn and make mistakes can affect policy shape. Learning agents having the ability to use similarity in their decision making is pivotal to the emergence of graduated-like sanctions in contexts where sanctioners do not redistribute collected fines. Given the fundamental nature similarity plays in learning and decision making in many intelligent creatures, it makes sense that allowing agents to use similarity in their reasoning to achieve long-run solutions for managing CPRs might produce results closer to what we observe in the real world - graduated sanctioning. Even in contexts where there is no similarity in reasoning, the best performing policy differs from what theory predicts. When agents learn and their exploration of the action space is intentional, our model shows modest additional fines over what theory predicts are required to discourage future exploration of actions which are particularly harmful to social welfare when fines are not redistributed. The importance of this finding, along with the role of similarity highlighted above, also suggests that modelers and analysts should perhaps be cautious about abstracting away from similarity and other learning processes when modelling behavior in other such contexts, as the policy recommendation may be fairly sensitive to these common cognitive processes.

The institutional design of sanctions can affect policy shape. If collections from sanctions do not feed back into the community, either because the fines cannot be redistributed or because of a lack of low-cost redistribution mechanisms, a social planner has incentive to keep sanctions relatively low. By contrast, the existence of effective reinvestment or redistribution opportunities that produce social welfare with the revenue from collected fines is sufficient to facilitate the long-run adoption of more draconian sanctions.

Social choice mechanisms can affect policy shape. Lastly, we observe that democracy can

solve the commons problem and does so with a fairly modest loss of social welfare when compared to the all knowing, benevolent social planner. The emergent policy shape is fairly unusual, however, resulting in excessive fining - in spite of the fact that representatives do not personally benefit from additional fines collected. While not explicitly explored, we suspect the extent to which social welfare is lost and the shape of the policy that emerges both depend highly on where the parties initially reside in the policy space and the fitness landscape of the policy space itself. This stems from the fact that winning representatives only need to outperform their next best performing rival and, in our model, do not take risks with their platforms when existing ones have proven successful.

6 Conclusion

As sustainability becomes more salient in the public consciousness, understanding when and under what conditions particular policies should be implemented and sustained to facilitate responsible use of common-pool resources grows ever more important. Adding to the literature on coordination and resource management spanning many fields, we have provided evidence for some sufficient conditions for the emergence of graduated (or draconian) sanctions as successful long-run policy solutions for managing CPRs. Additionally, as Ostrom had demonstrated in her work through the diversity of policy solutions she mentions were observed, we have started to develop a better understanding of the delicate relationship cognitive processes and policy constraints have with the types of policies that will prove most successful and how computational models can help us to pick at some elements of these relationships.

In a broader sense, we contribute to an ongoing discussion in the economics literature on the value of computational methods and where their applications in the field appropriately lie. A strength of agent-based models is their ability to allow researchers to explore worlds in which tractability assumptions can be relaxed. Further, the researcher can treat decision

making processes and model features as modular, substitutable components whose many combinations can be explored. With such methods in their tool-kits, researchers can begin to chip away at previously inaccessible regions of the research frontier, in tandem with utilization of more tried and true field methods to ground their findings. In our case, we use a model which encodes simple behavioral decision making rules and evolve policy solutions in a fairly unconstrained manner, but grounded in well studied theoretical models. This allows us to start bridging a gap between theory and what we observe in the natural world in a way that one method alone is incapable of.

Understanding that it is often easier to check if a policy solution is optimal than to find the optimal policy solution itself, tools that aim to automate the exploration of the policy space are of the utmost importance for solving complex policy problems. Such ideas are not new. For example, algorithmic game theory utilizes computational methods to solve practical real-time auction problems (Nisan et al., 2007). Still, we contribute to this literature in formulating one such way to apply computational exploration of policy questions.

While not the focus of this paper, this methodology may have applications for designing mechanisms which have desirable properties when faced by a wide variety of boundedly rational agents. Given the increasing prevalence of behavioral economics and recognition of humans' bounded rationality in decision theory, researchers may find value in tools like this one to find well-performing candidate policy solutions facing a variety of boundedly rational agent specifications. Perhaps someday such methods could be integrated into the early stages of policy solution exploration, after which small-scale studies can be performed to evaluate their performance in the wild.

Our model opens up future work. This framework could be used to investigate policies with dynamic sanctioning—that is, tracking individual agents' history of violating rules and fining them accordingly. It could also be used to investigate the role of agent heterogeneity in their ability to solve the commons problem. This model can also be used to investigate other political regimes; for example, variations on Competitive Direct Democracy in which

there are, for example, parties with ideological agendas along with their goal of maximizing votes, or the possibility of corruption or faithlessness. In this framework, future work could also explore conditions under which Ostrom's other design principles may emerge and how those principles relate to graduated sanctions and to each other.

References

A, Schlüter M Tavoni and Levin S, "The Survival of the Conformist: Social Pressure and Renewable Resource Management," *Journal of Theoretical Biology*, 2012, *299*, 152–61.

Andreoni, James and Laura K. Gee, "Gun for Hire: Delegated Enforcement and Peer Punishment in Public Goods Provision," *Journal of Public Economics*, 2012, *96* (11-12), 1036–1046.

Axtell, Robert, Robert Axelrod, Joshua M Epstein, and Michael D Cohen, "Aligning Simulation Models: A Case Study and Results," *Computational & Mathematical Organization Theory*, 1996, *1*, 123–141.

Baggio, Jacopo A, Allain J Barnett, Irene Perez-Ibara, Ute Brady, Elicia Ratajczyk, Nathan Rollins, Cathy Rubiños, Hoon C Shin, David J Yu, Rimjhim Aggarwal et al., "Explaining success and failure in the commons: the configural nature of Ostrom's institutional design principles," *International Journal of the Commons*, 2016, *10* (2), 417–439.

Bardhan, Pranab, "Analytics of the Institutions of Informal Cooperation in Rural Development," *World Development*, 1993, *21* (4), 633–639.

Bowles, Samuel and Jung-Kyoo Choi, "Coevolution of Farming and Private Property During the Early Holocene," *Proceedings of the National Academy of Sciences*, 2013, *110* (22), 8830–8835.

Boyd, Robert, Herbert Gintis, Samuel Bowles, and Peter J Richerson, "The Evolution of Altruistic Punishment," *Proceedings of the National Academy of Sciences*, 2003, *100* (6), 3531–3535.

Brown, John, "Some Tests of the Decay Theory of Immediate Memory," *Quarterly Journal of Experimental Psychology*, 1958, *10* (1), 12–21.

Camerer, Colin and Teck-Hua Ho, "Experience-Weighted Attraction Learning in Normal Form Games," *Econometrica*, 1999, *67* (4), 827–874.

Chaudhuri, Ananish, "Sustaining Cooperation in Laboratory Public Goods Experiments:

A Selective Survey of the Literature," *Experimental Economics*, 2011, *14*, 47–83.

Couto, Marta C, Jorge M Pacheco, and Francisco C Santos, "Governance of Risky Public Goods Under Graduated Punishment," *Journal of Theoretical Biology*, 2020, *505*, 110423.

De Geest, Lawrence R. and David C. Kingsley, "Norm Enforcement With Incomplete Information," *Journal of Economic Behavior & Organization*, 2021, *189*, 403–430.

_ **and John Miller**, "Using Social Choice to Solve Social Dilemmas," *Working Paper*, 2023.

Dubois, Didier, Lluís Godo, Henri Prade, and Adriana Zapico, "On the Possibilistic Decision Model: From Decision Under Uncertainty to Case-Based Decision," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 1999, *7* (06), 631–670.

Duffy, John, "Learning to Speculate: Experiments with Artificial and Real Agents," *Journal of Economic Dynamics and Control*, 2001, *25* (3-4), 295–319.

Eichberger, Jürgen and Ani Guerdjikova, "Case-Based Decision Theory: From the Choice of Actions to Reasoning About Theories," *Revue économique*, 2020, *71* (2), 283–306.

Engel, Christoph, "Social Preferences Can Make Imperfect Sanctions Work: Evidence From a Public Good Experiment," *Journal of Economic Behavior & Organization*, 2014, *108*, 343–353.

Erev, Ido and Alvin E. Roth, "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *The American Economic Review*, 1998, *88* (4), 848–881.

_ **and Alvin E Roth**, "Maximization, Learning, and Economic Behavior," *Proceedings of the National Academy of Sciences*, 2014, *111* (supplement_3), 10818–10825.

_ **, Eyal Ert, and Alvin E Roth**, "A Choice Prediction Competition for Market Entry Games: An Introduction," *Games*, 2010, *1* (2), 117–136.

Farmer, J Doyne and Shareen Joshi, "Price Dynamics of Common Trading Strategies,"

Journal of Economic Behavior & Organization, 2002, *49* (2), 149–171.

Fehr, Ernst and Simon Gächter, "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 2000, *90* (4), 980–994.

_ **and Tom Williams**, "Social Norms, Endogenous Sorting and the Culture of Cooperation," Working Paper 267, University of Zurich, Department of Economics 2018.

Franke, Reiner, "Reinforcement Learning in the El Farol Model," *Journal of Economic Behavior & Organization*, 2003, *51* (3), 367–388.

Gächter, Simon, Elke Renner, and Martin Sefton, "The Long-Run Benefits of Punishment," *Science*, 2008, *322*, 1510–1510.

Ghate, Rucha and Harini Nagendra, "Role of Monitoring in Institutional Performance: Forest Management in Maharashtra, India," *Conservation and Society*, 2005, *3* (2), 509–532.

Gilboa, Itzhak and David Schmeidler, "Case-Based Decision Theory," *The Quarterly Journal of Economics*, 1995, *110* (3), 605–639.

_ **and** _ , "Act Similarity in Case-Based Decision Theory," *Economic Theory*, 1997, *9*, 47–61.

_ **and** _ , *A Theory of Case-Based Decisions*, Cambridge University Press, 2001.

_ **,** _ **, and Peter P Wakker**, "Utility in Case-Based Decision Theory," *Journal of Economic Theory*, 2002, *105* (2), 483–502.

Guerrero, Omar A and Robert L Axtell, "Using Agentization for Exploring Firm and Labor Dynamics: A Methodological Tool for Theory Exploration and Validation," in "Emergent results of artificial economics," Springer, 2011, pp. 139–150.

Guilfoos, Todd and Andreas Duus Pape, "Predicting human cooperation in the Prisoner's Dilemma using case-based decision theory," *Theory and Decision*, 2016, *80*, 1–32.

Guttman, Norman and Harry I Kalish, "Discriminability and Stimulus Generalization," *Journal of Experimental Psychology*, 1956, *51* (1), 79.

Hume, David, *An Enquiry Concerning Human Understanding*, London, 1777.

Iwasa, Yoh and Joung-Hun Lee, "Graduated Punishment Is Efficient in Resource Management If People Are Heterogeneous," *Journal of Theoretical Biology*, 2013, *333*, 117–125.

Janssen, Marco and Elinor Ostrom, "Empirically Based, Agent-Based Models," *Ecology and Society*, 12 2006, *11*.

Jules, Selles, Bonhommeau Sylvain, Guillotreau Patrice, and Vallée Thomas, "Can the Threat of Economic Sanctions Ensure the Sustainability of International Fisheries? An Experiment of a Dynamic Non-cooperative CPR Game with Uncertain Tipping Point," *Environmental and Resource Economics*, 2020, *76* (1), 153–176.

Kinjo, Keita and Shinya Sugawara, "Predicting Empirical Patterns in Viewing Japanese TV Dramas Using Case-Based Decision Theory," *The BE Journal of Theoretical Economics*, 2016, *16* (2), 679–709.

Kirman, Alan P and Nicolaas J Vriend, "Evolving Market Structure: An ACE Model of Price Dispersion and Loyalty," *Journal of Economic Dynamics and Control*, 2001, *25* (3-4), 459–502.

LeBaron, Blake, "A Builder's Guide to Agent-Based Financial Markets," *Quantitative Finance*, 2001, *1* (2), 254–261.

Liu, Jia, Yohanes E. Riyanto, and Ruike Zhang, "Firing the Right Bullets: Exploring the Effectiveness of the Hired-Gun Mechanism in the Provision of Public Goods," *Journal of Economic Behavior & Organization*, 2020, *170*, 222–243.

Maja, Tavoni Alessandro Schlüter and Levin Simon, "Robustness of Norm-Driven Cooperation in the Commons," *Proc. R. Soc*, 2016, *283*.

Markussen, Thomas, Louis Putterman, and Jean-Robert Tyran, "Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes," *The Review of Economic Studies*, 2014, pp. 301–324.

Molleman, Lucas et al., "People Prefer Coordinated Punishment in Cooperative Interactions," *Nature Human Behaviour*, 2019, *3* (11), 1145–1153.

Moor, Tine De and Annelies Tukker, "Participation Versus Punishment: The Rela-

tionship Between Institutional Longevity and Sanctioning in the Early Modern Times," 2015.

_ , **Mike Farjam, René Van Weeren, Giangiacomo Bravo, Anders Forsman, Amineh Ghorbani, and Molood Ale Ebrahim Dehkordi**, "Taking Sanctioning Seriously: The Impact of Sanctions on the Resilience of Historical Commons in Europe," *Journal of Rural Studies*, 2021, *87*, 181–188.

Neugart, Michael, "Labor Market Policy Evaluation with ACE," *Journal of Economic Behavior & Organization*, 2008, *67* (2), 418–430.

Nevo, Ido and Ido Erev, "On Surprise, Change, and the Effect of Recent Outcomes," *Frontiers in Psychology*, 2012, *3*, 24.

Nicklisch, Alexander, Kristof Grechenig, and Christian Thöni, "Information-Sensitive Leviathans," *Journal of Public Economics*, 2016, *144*, 1–13.

Nisan, Noam, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani, *Algorithmic Game Theory*, New York, NY, USA: Cambridge University Press, 2007.

Oliveira, Fernando S, "The Emergence of Social Inequality: A Co-evolutionary Analysis," *Journal of Economic Behavior & Organization*, 2023, *215*, 192–206.

Ossadnik, Wolfgang, Dirk Wilmsmann, and Benedikt Niemann, "Experimental Evidence on Case-Based Decision Theory," *Theory and Decision*, 2013, *75*, 211–232.

Ostrom, Elinor, *Governing the Commons: The Evolution of Institutions for Collective Action.*, Cambridge University Press, 1990.

_ , "Design Principles in Long-Enduring Irrigation Institutions," *Water Resources Research*, 1993, *29* (7), 1907–1912.

_ , "Collective Action and the Evolution of Social Norms," *Journal of Economic Perspectives*, 2000, *14* (3), 137–158.

_ , "Do institutions for collective action evolve?," *Journal of Bioeconomics*, April 2014, *16* (1), 3–30.

_ , **James Walker, and Roy Gardner**, "Covenants With and Without a Sword: Self-

Governance Is Possible," *American Political Science Review*, 1992, *86* (2), 404–417.

_ **, Roy Gardner, and James Walker**, *Rules, games, and common-pool resources*, University of Michigan press, 1994.

Pape, Andreas Duus and Kenneth J Kurtz, "Evaluating Case-Based Decision Theory: Predicting Empirical Patterns of Human Classification Learning," *Games and Economic Behavior*, 2013, *82*, 52–65.

Pigou, Arthur Cecil, *A study in public finance*, Macmillan, 1929.

Radoc, Benjamin, "Case-Based Investing: Stock Selection Under Uncertainty," *Journal of Behavioral and Experimental Finance*, 2018, *17*, 53–59.

Raihani, Nichola J., Alex Thornton, and Redouan Bshary, "Punishment and Cooperation in Nature," *Trends in Ecology & Evolution*, 2012, *27*, 288–295.

Rubinos, Cathy, "Commons Governance for Robust Systems: Irrigation Systems Study Under a Multi-Method Approach," *Doctoral dissertation - Arizona State University*, 2017.

Sarin, Rajiv and Farshid Vahid, "Predicting How People Play Games: A Simple Dynamic Model of Choice," *Games and Economic Behavior*, 2001, *34* (1), 104–122.

Selten, Reinhard, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 1975, *4*, 25–55.

_ **and Thorsten Chmura**, "Stationary Concepts for Experimental 2×2-Games," *American Economic Review*, 2008, *98* (3), 938–966.

Sethi, Rajiv and E. Somanathan, "The Evolution of Social Norms in Common Property Resource Use.," *The American Economic Review*, 1996, *86* (4), 766–788.

Shepard, R.N., "Toward a Universal Law of Generalization for Psychological Science," *Science*, 1987, *237* (4820), 1317.

Shin, Hoon C, J Yu David, Samuel Park, John M Anderies, Joshua K Abbott, Marco A Janssen, and TK Ahn, "How Do Resource Mobility and Group Size Affect Institutional Arrangements for Rule Enforcement? A Qualitative Comparative Analysis of Fishing Groups in South Korea," *Ecological Economics*, 2020, *174*, 106657.

Skinner, Burrhus Frederic, "The Experimental Analysis of Behavior," *American Scientist*, 1957, *45* (4), 343–371.

Thomas, Priya and Todd Guilfoos, "Case-Based Reasoning and Dynamic Choice Modeling," *Land Economics*, 2023, *99* (1), 103–121.

Traulsen, Arne and Martin Nowak, "Evolution of Cooperation by Multilevel Selection," *Proceedings of the National Academy of Sciences*, 2006, *103* (29), 10952–10955.

Tsatsoulis, Costas, Qing Cheng, and H-Y Wei, "Integrating Case-Based Reasoning and Decision Theory," *IEEE Expert*, 1997, *12* (4), 46–55.

van Klingeren, Fijnanda and Vincent Buskens, "Graduated Sanctioning, Endogenous Institutions and Sustainable Cooperation in Common-Pool Resources: An Experimental Test," *Rationality and Society*, 2024, *36* (2), 183–229.

Visser, Martine and Justine Burns, "Inequality, Social Sanctions and Cooperation Within South African Fishing Communities," *Journal of Economic Behavior & Organization*, 2015, *118*, 95–109.

Waring, Timothy M, Sandra H Goff, and Paul E Smaldino, "The Coevolution of Economic Institutions and Sustainable Consumption Via Cultural Group Selection," *Ecological Economics*, 2017, *131*, 524–532.

Wilson, David Sloan, Elinor Ostrom, and Michael E Cox, "Generalizing the Core Design Principles for the Efficacy of Groups," *Journal of Economic Behavior & Organization*, 2013, *90*, S21–S32.

Appendices

Solving the Theoretical Harvest Game

Rational, Selfish Agents Facing No Policy:

Agents try to choose h_i to maximize their own payoff given by Equation 3, hence their maximization problem is

$$\max_{h_i \in [0,H]} [h_i - \alpha \overline{h} - \beta \overline{h}^2] \tag{A1}$$

Where $\overline{h}$ is the average harvest choice given in Equation 2. Taking first order conditions, we get

$$1 - \alpha \frac{1}{N} - \frac{2\beta}{N^2}(h_i^* + \Sigma_{-i} h_j) = 0 \tag{A2}$$

By symmetry, we can simplify the latter portion of the equation in the following way

$$h_i^* + \Sigma_{-i} h_j = N h_i^* \tag{A3}$$

Now we can substitute Equation A3 into Equation A2 and do some simplification.

$$1 - \alpha \frac{1}{N} - \frac{2\beta}{N^2} N h_i^* = 0 \tag{A4}$$

$$1 - \alpha \frac{1}{N} = \frac{2\beta}{N^2} N h_i^* \tag{A5}$$

$$1 - \alpha\frac{1}{N} = \frac{2\beta}{N}h_i^* \tag{A6}$$

$$\frac{N}{2\beta}(1 - \alpha\frac{1}{N}) = h_i^* \tag{A7}$$

Simplifying once more we find the competitive equilibrium solution as included above.

$$h_{CE} = \frac{N - \alpha}{2\beta} \tag{4}$$

Rational, Altruistic Agents Facing No Policy:

In the altruistic agents' maximization problems, h_i is chosen to maximize the sum of all players' payoffs. This problem is well known to be equivalent to the benevolent social planner's problem in which they must choose the harvest levels for the agents to maximize the sum of all player's payoffs. This is given by

$$\max_{\{h_1,\ldots,h_N|\forall h_i, h_i\in[0,H]\}} \Sigma_{j=1}^{N}[h_j - \alpha\overline{h} - \beta\overline{h}^2] \tag{A4}$$

Again with $\overline{h}$ as the average harvest choice given in Equation 2. By symmetry, the problem can be reduced to

$$\max_{h\in[0,H]} \Sigma_{j=1}^{N}[h - \alpha h - \beta h^2] \tag{A5}$$

$$\max_{h\in[0,H]} N[h - \alpha h - \beta h^2] \tag{A6}$$

Taking first order conditions, we get

$$1 - \alpha - 2\beta h^* = 0 \tag{A7}$$

$$1 - \alpha = 2\beta h^* \tag{A8}$$

As included above, we find

$$h_{SO} = \frac{1 - \alpha}{2\beta} \tag{A9}$$

Rational, Benevolent, Fully Informed Social Planner Choosing Policy:

In this problem, the social planner does not have direct control over agent harvest levels. Instead, the planner must choose a policy f(.) to influence the choices selfish agents make, aiming to maximize social welfare. First, let us assume the planner aims to maximize social welfare excluding the penalty to agent benefit incurred from the policy itself. Later, we will show this distinction is of little consequence in the theoretical model. Thus the social planner must choose f(.) to solve the following maximization problem

$$\max_{f(.)} \Sigma_{j=1}^{N} \pi_i(h_j) \tag{A10}$$

Knowing that agents will choose hi conditional on what f(.) is

$$\pi_i(h_i, f(h_i)) = h_i - \alpha\overline{h} - \beta\overline{h}^2 - f(h_i) \tag{5}$$

Then it should be clear, by backwards induction, the social planner needs to pick a policy

function f(.) such that agents, when facing the policy function, choose the socially optimal level of h_i given in Equation **??**. A policy function f(.) will induce this if the following condition holds

$$\forall h_i \neq h_{SO}, \pi_i(h_{SO}, f(h_{SO})|h_j = h_{SO} \forall j \neq i) \geq \pi_i(h_i, f(h_i)|h_j = h_{SO} \forall j \neq i) \tag{A11}$$

That is, the payoff of choosing the socially optimal harvest level h_{SO} when everyone else is also choosing the socially optimal harvest level has to be at least as good as choosing anything else. Utilizing the fact that the policy function enters agent payoffs additively, we can create the following equivalent inequality which must hold for our policy function f(.) to maximize social welfare:

$$f(h_i) \geq \pi_i(h_i|h_j = h_{SO} \forall j \neq i) - [\pi_i(h_{SO}|h_j = h_{SO} \forall j \neq i) - f(h_{SO})] \tag{A12}$$

Simply put, agents have to be penalized at least the marginal benefit they would get if they choose something other than h_{SO}.

Using Equation 3 we can rewrite part of Equation A12 as the following

$$\pi_i(h_i|h_j = h_{SO} \forall j \neq i) = h_i - \alpha \frac{1}{N} \Sigma_{j=1}^{N} h_j - \beta (\frac{1}{N} \Sigma_{j=1}^{N} h_j)^2 \tag{A13}$$

which simplifies to

$$\pi_i(h_i|h_j = h_{SO} \forall j \neq i) = h_i - \alpha \frac{1}{N} (h_i + (N-1)h_{SO}) - \beta (\frac{1}{N} (h_i + (N-1)h_{SO}))^2 \tag{A14}$$

Substituting Equation A14 into Equation A12 we get

$$f(h_i) \geq (h_i - \alpha \frac{1}{N} (h_i + (N-1)h_{SO}) - \beta (\frac{1}{N} (h_i + (N-1)h_{SO}))^2)$$

$$- [(1 - \alpha - \beta(h_{SO}))h_{SO} - f(h_{SO})] \tag{A15}$$

which can be rewritten after quite a bit of algebra as

$$f(h_i) = \begin{cases} X & \text{if } h_i = h_{SO} \\ j \in [X + A + Bh_i + Ch_i^2, \text{inf}) & \text{otherwise} \end{cases} \tag{A16}$$

where

$$A = \frac{\beta(2N-1)}{N^2} h_{SO}^2 - (1 - \frac{\alpha}{N}) h_{SO}$$

$$B = 1 - \frac{\alpha}{N} - \frac{2\beta(N-1)}{N^2} h_{SO}$$

$$C = \frac{-\beta}{N^2}$$

and x is whatever the social planner chooses to fine or punish an agent who chooses the socially optimal level of harvest. Any policy that satisfies Equation A16 is optimal for this social planner.

In our context, we impose one additional constraint. Since we do not allow negative fines (subsidies or rewards), the lowest possible fine allowed is 0. Incorporating this into Equation A16, we get the more constrained set of possible policy solutions below.

$$f(h_i) = \begin{cases} X & \text{if } h_i = h_{SO} \\ j \in [\max(0, X + A + Bh_i + Ch_i^2), \text{inf}) & \text{otherwise} \end{cases} \tag{A17}$$

where

$$A = \frac{\beta(2N-1)}{N^2} h_{SO}^2 - (1 - \frac{\alpha}{N}) h_{SO}$$

$$B = 1 - \frac{\alpha}{N} - \frac{2\beta(N-1)}{N^2} h_{SO}$$

$$C = \frac{-\beta}{N^2}$$

Now if we suppose the social planner wants to maximize social welfare, considering f(.) as harmful to social welfare as one might conventionally think would be the case, the solution set given in Equation A16 is only altered such that X = 0 must hold. This is because rational agents will only ever incur the penalties associated with on equilibrium path actions. Since the only on equilibrium path penalty the players face is f(h_{SO}), the negative effect fines have on social welfare is minimized by simply making the fine associated with choosing $f(h_{SO}) = 0$. Plugging this condition into Equation A17, we get the policy solution in Equation 6. As we will show in the next section, the solution set remains unchanged in the trembling hand version of the problem when fines are redistributed.

Solving the Harvest Game for Rational Agents with Trembling Hands:

Perfectly Informed Rational Agents with Mistake-Making

There are many Nash equilibria without imposing trembling hands, as seen above. The reason, intuitively, is: consider a sanctioning strategy (a fine vector) where agents are charged fines for non-socially optimal harvest levels which are high enough to make the payoff associated with choosing h_{SO} either the highest or tied for the highest. For any such policy and rational agents, you can increase the fines associated with any or all non-socially optimal choices without affecting behavior or social welfare. Behavior is unaffected because the increase in off-equilibrium fines simply makes undesirable choices less desirable for rational agents. They will continue to choose h_{SO} and not choose other actions. Social welfare is unaffected because behavior is unaffected and the payoff associated with equilibrium behavior has not changed. The differences in the payoffs of off-equilibrium choices are irrelevant for the social welfare of rational agents because they never incur these fines anyway. In our computational models, our agents explore, make mistakes, and learn from those experiences, so off-equilibrium fines will play a real role in social welfare generated under a policy. To

bring the theoretical model one step closer to our computational models, we extend the theoretical model presented above by allowing rational agents to make mistakes via trembling hands.

In any game with agents who have trembling hands, we imagine all players, after choosing their strategy, have a small probability ε of "making a mistake." When a mistake is made, the player ignores their intended action and instead chooses an action from the action set. In the Harvest Game, this means agents who choose their harvest level have a small chance ε to harvest a randomly chosen level instead. For simplicity, we first imagine errors are drawn uniformly from the action set, though we will discuss later the implications of local error making.

The game is solved for an arbitrary ε, and then ε is taken to 0. Note that as ε is taken to 0, behavior completely coincides with Nash reasoners. We compare these implied behaviors to the behaviors exhibited by our boundedly rational learning agents in our computational model under the Private Provisions Regime, described in Section 3.4.1.

Social welfare is now affected by off-equilibrium fines, as the social planner solves the problem as if agents will mistakenly choose off-equilibrium choices with some probability ε. When fines are not redistributed, the trembling hand solution reduces the set of possible equilibria found to a single solution - the minimum fine in the solution set without trembling hand agents (See Equation 7). For the parameters given here, , this means the unique policy solution is given by This reduction of the policy solution set is due to the policy maker recognizing that having higher fines than necessary to correct behavior comes at an additional welfare cost to agents ε of the time.

When the social planner utilizes fines with redistribution (such that there is no net loss to social welfare), the set of policy solutions completely coincides with the set of solutions for rational agents without trembling hands given in Equation 6. Note this extends directly from the fact that while fines are painful to the agents that face them, redistribution of those fines ensures there is no net loss of social welfare. This means fining any amount is equivalent,

so long as it is sufficiently high to deter agents from choosing something other than the socially optimal level.[7] We compare these policy solutions to the policy solutions found in our computational models with policy, The Social Planner Regime and The Competitive Direct Democracy Regime, in Sections 4.2 and 4.3.

We felt a trembling hand equilibrium solution was a closer theoretical model to compare our computational model to since both the policy that falls out of trembling hand equilibrium and our computational model take into account the agents playing some non-equilibrium path actions in their policy solution. It is also true that error making by players in the trembling hand problem goes to 0, but so does individual agent exploration in our computational model, as the exploration rate goes to 0, further aligning these models.

It can also be shown that if trembling-hand agents make errors locally instead of uniformly, they also fail to refine the solution set to graduated sanctions and can even exclude them from the solution set in the same way that the policy solution for trembling hand agents with uniform errors does. For fines without redistribution, regardless of the distribution of errors used, we find the following:

1. If there is a non-zero chance a particular action $\hat{h}$ can be chosen by mistake by an agent intending to choose the socially optimal level h_{SO}, the fine associated with $\hat{h}$ coincides with its size given in the trembling hand problem solution with uniform errors.

2. If there is no possible chance that a particular action $\hat{h}$ could be chosen by mistake by someone intending to play h_{SO}, then the fine associated with $\hat{h}$ can coincide with any of the possible levels given in the theoretical policy solution for agents without trembling hands. That is, fines can be arbitrarily higher than the minimum corrective level without any effect on social welfare.

Note that on one extreme, where agents have some chance to erroneously pick any action given intended socially optimal behavior, this solution completely coincides with the trembling hand with uniform errors. On the other extreme, as the set of local actions agents

[7]For more details, see Appendix 6.

can end up choosing by mistake shrinks, we approach the original Nash solution without mistake-making.

In the case imposing fines with redistribution, the solution set once again remains unchanged.

We find that these theoretical models with and without mistake-making are insufficient to explain graduated sanctions in contexts both with and without redistribution. We can see the solution found for fine-based policies without redistribution is not graduated when rational agents have trembling hands (as will be made clear in the next section) and the sets of policy solutions in all other contexts explored indicate that part or all of the policy shape is irrelevant so long as it lies above the lower bound. This does not align with what Ostrom observed (as will be discussed in the next section).

In trembling hand equilibrium, agents will still choose h_i to maximize their own payoff. Agents in this context are distinct from in the typical context in that they have a small chance, ε, to forgo playing their intended harvest level h_i. Instead, a random harvest level is chosen uniformly from the action set, which in our case, is the interval [0, H] (recall H is the maximum harvest level possible. This dynamic is meant to capture mistake-making (e.g. a 'mouse slip'). We can find a trembling hand equilibrium by solving the game for an arbitrary (but small) ε, and then taking the limit as ε approaches 0.

Thus, the selfish agent's problem can be reformulated as:

$$\max_{\{\hat{h_1},\ldots,\hat{h_N}|\forall\hat{h_i},\hat{h_i}\in[0,H]\}} \hat{h_j} - \alpha\overline{\hat{h}} - \beta\overline{\hat{h}}^2 \tag{A18}$$

with

$$\hat{h_k} = \begin{cases} h_k & \text{with probability } 1-\varepsilon \\ h_{random}\ U[0,H] & \text{otherwise} \end{cases} \tag{A19}$$

Since all trembling hand equilibrium are Nash equilibrium and there always exists a

trembling hand equilibrium, and given that this problem has a unique (symmetric) solution, we can see that that the selfish agent's optimal harvest level remains unchanged from the general case, as seen in Equation 4.

Similarly, an altruistic agent's problem is given by

$$\max_{\{\hat{h_1},\ldots,\hat{h_N}|\forall\hat{h_i},\hat{h_i}\in[0,H]\}} \Sigma_{j=1}^{N}[\hat{h_j} - \alpha\overline{\hat{h}} - \beta\overline{\hat{h}}^2] \tag{A20}$$

and once again, we find that the altruistic agent's symmetric Nash equilibrium remains unchanged (given in Equation 5) from the base case as ε approaches 0 as, once again, there exists a unique symmetric Nash equilibrium.

For a policy maker facing trembling agents without redistribution, their problem statement is given as follows:

$$\max_{f(.)} \Sigma_{j=1}^{N}\pi_i(\hat{h_k}) \tag{A21}$$

i.e.

$$\max_{f(.)} \Sigma_{j=1}^{N}[\hat{h_j} - \alpha\overline{\hat{h}} - \beta\overline{\hat{h}}^2 - f(\hat{h})] \tag{A22}$$

The policy maker's solution set (from Equation 6) collapses to a single solution, given by:

$$f(h_i) = \begin{cases} 0 & \text{if } h_i = h_{SO} \\ \max(0, A + Bh_i + Ch_i^2) & \text{otherwise} \end{cases} \tag{A23}$$

where

$$A = \frac{\beta(2N-1)}{N^2}h_{SO}^2 - (1 - \frac{\alpha}{N})h_{SO}$$

$$B = 1 - \frac{\alpha}{N} - \frac{2\beta(N-1)}{N^2} h_{SO}$$

$$C = \frac{-\beta}{N^2}$$

This is simply the floor of the Nash equilibrium solution set with X = 0. This solution refinement is simply the result of the fact that the policy maker believes there is some chance that any harvest level will be chosen by an agent. Given this, the off equilibrium path harvest levels can not have arbitrarily high punishments in equilibrium. Instead, the social planner fines non-socially optimal behavior just enough to make players indifferent (between h_{SO} and any alternatives), which minimizes the loss to social welfare incurred by agents who happen to tremble.

In the social planner's problem with redistribution, however, the solution remains unchanged from the Nash equilibrium solution set given in Equation 6. Since punishment does not affect net social welfare (outside of how it steers agent behavior), incurring a higher than need-be off equilibrium fine remains irrelevant for social welfare.

Agent Decision Making

At the start of the game, Agents:

- Initialize

Each round of the game, agents:

1. Choose an action
2. Update their action scores
3. Update their exploration rate

Initialization

Each agent i starts with an action set and a vector of scores S associated with each action in that action set. An action's score is simply the agent's own running average payoff from past experiences using an action. Since at time of initialization no such experiences have been accrued yet, each action in the agent's action set is initially assigned some arbitrarily large score Z which will be replaced with with the running average performance after a single experience has been accrued. In our case, agent i chooses $h_i \in \{0, ..., H\}$, so we can write our initialization step as follows

$$S_i(h_i) = Z \quad \forall h_i \in \{0, ..., H\} \tag{B1}$$

Agents also keep track of how often they have chosen each action with a vector freq. At the start of the model, each entry in freq is 0, as no action has been taken yet. Formally

$$freq_{i,t=0}(h_i) = 0 \quad \forall h_i \in \{0, ..., H\} \tag{B2}$$

Each agent also starts with a few initial parameters which will guide their rate of exploration. Agents have an initial probability to explore p_t and a rate at which that exploration rate decays λ. We set $p_t = 1$ and $\lambda = 0.0005$ as our baseline values.

1. Choosing an Action:

First agents must decide whether to explore or not. Agents have a probability of p_t to **explore** and 1-p_t to **exploit**.

Explore

The agent chooses an action from your action set randomly. The probability with which an agent chooses an action is proportional to its score.

$$Prob(h_i) = \frac{S_i(h_i)}{\Sigma_{j=0}^{H} h_j} \tag{B3}$$

Exploit

The agent chooses the action with the highest score.

$$h_i = \arg\max_{h_i} S_i(h_i) \tag{B4}$$

Note that in the case of a tied score, an action is chosen randomly from the tied candidates with equal probability.

2. Updating Action Scores:

First, when an action is chosen, we must update the frequency vector $freq_{i,t}$ to reflect the agent chose h_i this round.

$$freq_{i,t}(\hat{h_i}) = \begin{cases} freq_{i,t-1}(\hat{h_i}) + 1 & \text{if } \hat{h_i} = h_i \\ freq_{i,t-1}(\hat{h_i}) & \text{otherwise} \end{cases} \tag{B5}$$

Next, we need to update the score associated with each of our actions. Recall at initialization, each action has some high level of attraction. Three such cases arise during this step.

1. For actions not chosen this round, their scores remain unchanged.

2. If this is the first time the agent has chosen h_i (ie. if $freq_{i,t-1} = 0$), we replace the score which was set during initialization with the normalized performance of the action this period.

3. If the agent has played h_i before, the agent updates the score as a running average of all past normalized payoffs observed playing the action.

All this is to say, if this is the first time an action is chosen, the payoff received replaces the initial score Z with which the action was initialized. If the action has been chosen before, we update the running average performance with the payoff received instead.

This is formalized as

$$S_{i,t}(\hat{h_i}) = \begin{cases} \pi_i(\hat{h_i}) & \text{if } \hat{h_i} = h_i \text{ and } freq_{i,t-1} = 0 \\ \frac{1}{freq_{i,t}(h_i)}\pi_i(\hat{h_i}) + \frac{freq_{i,t-1}(h_i)}{freq_{i,t}(h_i)} S_{i,t-1}(h_i) & \text{if } \hat{h_i} = h_i \text{ and } freq_{i,t-1} > 0 \\ S_{i,t-1}(h_i) & \text{otherwise} \end{cases} \tag{B6}$$

where

$$\hat{\pi_i}(h_i) = \pi_i(h_i) - \min[\pi_i(h_i)] \tag{B7}$$

Intuitively, the linear transformation of utility performed to construct $\hat{\pi_i}(h_i)$ ensures non-negative scoring, which is required for how we perform exploration in Equation B3. Importantly, this shift in utility is subtracted back out when doing social welfare comparisons to ensure the payoffs accrued in both our theoretical and computational models remain comparable.

When agents utilize similarity, they also update nearby actions as well. In the case of our similarity leveraging agents, an action is only considered similar if it is adjacent to the chosen action in the discretized action space. Mechanically, the equations given above are updated as follows

$$S_{i,t}(\hat{h_i}) = \begin{cases} \hat{\pi_i}(h_i) & \text{if } |\hat{h_i} - h_i| \leq 1 \text{ and } freq_{i,t-1} = 0 \\ \frac{1}{freq_{i,t}(h_i)}\hat{\pi_i}(h_i) + \frac{freq_{i,t-1}(h_i)}{freq_{i,t}(h_i)} S_{i,t-1}(h_i) & \text{if } |\hat{h_i} - h_i| \leq 1 \text{ and } freq_{i,t-1} > 0 \\ S_{i,t-1}(h_i) & \text{otherwise} \end{cases} \tag{B8}$$

with

$$freq_{i,t}(\hat{h_i}) = \begin{cases} freq_{i,t-1}(\hat{h_i}) + 1 & \text{if } \hat{h_i} = h_i \\ freq_{i,t-1}(\hat{h_i}) + \frac{1}{e} & \text{if } |\hat{h_i} - h_i| = 1 \\ freq_{i,t-1}(\hat{h_i}) & \text{otherwise} \end{cases} \tag{B9}$$

and

$$\hat{\pi_i}(h_i) = \pi_i(h_i) - \min[\pi_i(h_i)] \tag{B10}$$

Intuitively, actions chosen are updated as before, but now nearby actions have their attraction (the expected payoff of the action) updated as well. It is updated to a lesser degree $\frac{1}{e}$.

3. Updating Exploration Rate:

The exploration rate p is updated using the decay rate λ in the following way:

$$p_{t+1} = p_t e^{-\lambda} \tag{B11}$$

Social Planner Decision Making

At the start of the simulation, the Social Planner:

- Initializes

Each round of the simulation, the social planner:

1. Creates candidate policies
2. Evaluates the candidate policies
3. Stores the best candidate policy

Presently each simulation is run for 10,000 iterations and we repeat the simulation 20 times, each with a new initialization. The policy which performed best across all 20 simulations is the social planner's best found solution to the commons problem.

Initialization

The social planner starts with fine vector $f_{t=0}(h_i)$ with an entry for each action in the agents' action set representing the penalty agents will get for choosing that action. Each entry in the vector is independently and identically drawn from Uniform[0, M], where M is the maximum fine allowable. In our case, agent i chooses $h_i \in 0, ..., H$, so we can write our initialization step as

$$f_{t=0}(h_i) = \theta \sim Uniform[0, M] \qquad \forall h_i \in \{0, ..., H\} \tag{C1}$$

The social planner also starts with a few initial parameters which will guide how they explore the policy space, q_{mutate} and q_{range} which we will discuss shortly. As a baseline we select $q_{mutate} = 0.5$ and a $q_{range} = 0.1$

1. Creating Candidate Policies:

From the policy which performed best last round, $f_{t-1}(h_i)$, the social planner creates R candidate policies $\{\widehat{f_{t,r=1}(h_i)}, ..., \widehat{f_{t,r=R}(h_i)}\}$ to consider, which are variations of $f_{t-1}(h_i)$.

To construct a candidate policy, first a copy of $f_{t-1}(h_i)$ Then each dimension of the policy (each position in the vector) has a q_{mutate} chance to have random noise added to it. This random noise is drawn from a normal distribution with a mean of 0 and a variance scaled by our q_{range} parameter. We can formalize the construction of the kth candidate policy as follows:

$$\widehat{f_{t,r=k}(h)} = \begin{cases} f_{t-1}(h) & \text{with prob } 1 - q_{mutate} \\ f_{t-1}(h) + z & \text{otherwise} \end{cases} \qquad \forall h_i \in \{0, ..., H\} \tag{C2}$$

where

$$z \sim \text{Normal}(0, M * q_{range}) \tag{C3}$$

and M is the maximum fine allowable.

It is sometimes the case that after a dimension has random noise added to it, it falls outside of the allowable range for policy values [0, M]. In such cases, we will replace this illegal policy dimension specification which we will denote as d for now with a new value drawn in the following way:

$$\widehat{f_{t,r=k}(h)}|d \notin [0, M] = \begin{cases} y \sim Uniform[0, f_{t,r=k}(h)] & \text{if } d < 0 \\ y \sim Uniform[f_{t,r=k}(h), M] & \text{otherwise} \end{cases} \tag{C4}$$

Implementing it this way guarantees an allowable value by the second draw and combats potential directional biases on policy refinement when values in the fine vector are close to the allowable boundary.

2. Evaluating Candidate Policies:

The best performing policy from last round $f_{t-1}(h_i)$ and the R candidate policies $\widehat{f_{t,r=1}(h_i)}, ..., \widehat{f_{t,r=R}(h_i)}$ are all evaluated this round. To do this, the social planner runs the repeated game under each policy a number of times and then collects the average social welfare accrued across runs when the agents faced the policy in question. Thus, a vector of $[\overline{\Psi}(f_{t-1}(h_i)), \overline{\Psi}(f_{t,r=1}(h_i)), ..., \overline{\Psi}(f_{t,r=R}(h_i))]$. Our results come from a social planner who constructs 7 new candidates each round of the simulation (R=7), each of which is run 5 times to forecast the average social welfare the policy is expected to produce.

3. Storing the Best Policy:

Finally, the social planner compares the average social welfare generated by the R candidate policies against the last round's best performer. The one expected to produce the highest social welfare is stored as the best performer of this round. Formally

$$f_t(h_i) = \underset{f \in C}{\operatorname{argmax}}\, \overline{\Psi}(f) \tag{C5}$$

where

$$C = \{f_{t-1}(h_i), \widehat{f_{t,r=1}(h_i)}, ..., \widehat{f_{t,r=R}(h_i)}\} \tag{C6}$$

Social Choice via Democracy

At the start of the simulation, our democracy module:

- Initializes

Each round of the simulation, an election cycle occurs in the following way:

1. Agents forecast well-being under policies
2. Agents vote for a policy
3. Representatives update their platforms

The simulation runs for 20,000 rounds (election cycles). The policy which is adopted at the end of this process is considered the long run policy solution with which the agents aim to solve the social dilemma with.

Initialization

The model starts with N representatives, each of which will be given their own initial platform (policy solution) in much the same way the social planner received their.

$$f_{t=0,n=l}(h_i) = \delta \sim U[0, M] \qquad \forall h_i \in \{0, ..., H\} \tag{D1}$$

where n denotes the representative's id.

In a similar fashion to the social planner, global values are also set for how policy solutions are to be explored, utilizing the same parameters from before: q_{mutate} and q_{range}. Again, as a baseline we select $q_{mutate} = 0.5$ and a $q_{range} = 0.1$. Additionally we must choose a number of parties N. We investigate the simple case of N=2.

1. Agents Forecast Well-Being Under Policies:

For each platform a representative has proposed, agents run forecasts of their utility under the policy $\widehat{\pi}_{i,t}(f_{t,n=l}(.))$ by facing the policy a number of times in their head and then calculating how well they do on average. If the policy is incumbent, they add their forecasts to the policies past performance. Formally

$$\widehat{\pi}_{i,t}(f_{t,n=l}(.)) = \begin{cases} \frac{1}{2}[\overline{\pi}_{i,t}(h_i, f_{t,n=l}(h_i) + \overline{\pi}_{i,t-1}(h_i, f_{t-1,n=l}(h_i))] & \text{if } f_{t,n=l}(h_i)) = f_{t-1,n=l}(h_i)) \\ \overline{\pi}_{i,t}(h_i, f_{t,n=l}(h_i)) & \text{otherwise} \end{cases} \tag{D2}$$

From this, each agent produces a vector of welfare forecasts, one for each platform, denoted Q = $\{\widehat{\pi}_{i,t}(f_{t,n=1}(.))$, ..., $\widehat{\pi}_{i,t}(f_{t,n=N}(.))\}$

2. Agents Vote for a Policy:

Now having forecasted how well each agent expects each policy to perform, the agents vote for the policies which they believe will give them the most utility on average. Formally

$$f_t(h_i) = \underset{f \in Q}{\text{argmax}} \, \overline{\widehat{\pi}_{i,t}(f)} \tag{D3}$$

The policy which wins the most votes is implemented, with ties broken randomly. We denote the boolean indicating if a representative won majority vote as $w(f_{t,n=l})$.

3. Representatives Update their Platforms:

After the election, the winning representative makes no change to their policy while all representatives who lost the election reconsider their strategy. First, they see if their plat-

form from last round was able to attract more votes $v(\widehat{f}_{t,n=l}(.))$ than in the previous round $v(\widehat{f}_{t,n=l}(.))$. The better performer is saved as the representatives baseline platform, with ties going to the most recent policy. This is given formally below as

$$f_{t,n=l}(.) = \begin{cases} \widehat{f}_{t,n=l}(h_i)) & \text{if } w(f_{t,n=l}(h_i)) = 0 \text{ and } v(\widehat{f}_{t,n=l}(h_i)) \geq v(\widehat{f}_{t-1,n=l}(h_i)) \\ f_{t-1,n=l}(h_i) & \text{otherwise} \end{cases} \tag{D4}$$

Next, the representative decide what platform to run on for the next election cycle. For the winner, this is easy as they run on their core platform $f_{t,n=l}$. For the losers of this cycle, they instead try a deviation from their baseline platform. This variant platform is created in much the same way as the social planner produces a candidate. Formally

$$\widehat{f_{t+1,n=l}}(h) = \begin{cases} f_{t-1,n=l}(h) + z & \text{if } w(f_{t,n=l}(h_i)) = 0 \text{ with prob } q_{mutate} \\ f_{t,n=l}(h) & \text{otherwise} \end{cases} \qquad \forall h_i \in \{0, ..., H\} \tag{D5}$$

where

$$z \sim \text{Normal}(0, M * q_{range}) \tag{D6}$$