Evolving Sustainable Institutions in Agent-Based Simulations with Learning

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

June 27, 2024

Abstract

Elinor Ostrom identified eight design principles for the management of common-pool resources across hundreds of case studies. We develop a novel computational model in which learning agents intentionally explore the action space in a common resource game under different policy regimes to test the conditions in which one of Ostrom's design principles, graduated sanctions, emerges. We characterize the long-run policies that emerge top-down via a computational social planner and bottom-up via democracy, modelled as an endogenous self-governance process.

First, we find that graduated sanctions emerge top-down via a social planner who utilizes a fine-based policy without redistribution, but only when agents utilize similarity in their decision-making process. Next, we find that, when policy makers are able to redistribute fines, draconian style sanctions emerge. We also demonstrate that implementing the theoretical solution for rational agents who fully understand the game can forgo substantial potential gains in social welfare. Finally, we observe that, when agents participate in "democracy" (a bottom-up policy selection mechanism via voting for representatives) they are able to solve the commons problem fairly well, though we do not observe graduated sanction emerge in this context.

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

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

[*] 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

[§] Department of Law, Northwestern University, United States of America

1 Introduction

In Elinor Ostrom's seminal work (Ostrom, 1990), she compiled hundreds of case studies of communities that successfully managed common-pool resources (CPRs). Across these communities, which 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, indicating these design principles were (in some cases) discovered independently. Ostrom points out traditional schools of economic thought do not appear prepared to explain this empirical regularity in how common resources are managed in the world (Ostrom, 1990). This work forms the basis of a collective action theory where individuals form institutions through self-organization.

Ostrom's Institutional Analysis and Development (IAD) framework posits the design principles can emerge through a deliberative process by agents set in an environment who devise their own policy solutions. Further, she suggests this complex process cannot be modeled as a simple game as one might see in a game theoretic model, instead suggesting 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). If these principles tend to improve management of CPRs, then it's possible to see the emergence of these principles in an appropriately-defined, adaptive/evolutionary agent-based model in which agents interact with a CPR. 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 focus on the the emergence of the principle of graduated sanctions, to see whether and how it might emerge in this context. We approach the problem iteratively by building out our model in 3 distinct phases and observing model behavior during each. [1] In the Private Provision Model, learning agents

[1] The code for this project is publicly available and can be found at `https://github.com/chriszosh1/EvolvingSustainableInstitutions`

play the CPR game facing no policy. In the Social Planner Model, the learning agents play the same game facing exogenously given policy via a benevolent social planner whom is experimenting with policies to improve social welfare. Finally in the Democracy Model, the learning agents play the game once more, but now participate in forming the policy they'll face via voting in a democracy.

We find 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 via voting, they are able to solve the commons problem, though sub-optimally through excessive fining. We also delve deeper into why both the theoretical solution and democracy achieve sub-optimal levels of social welfare.

An overview of the structure of the remaining paper is given as follows. In Section 2 we detail related works and where our contribution fits in. In Section 3, we describe the underlying theoretical model and describe the shape of the policy solution found in the theoretical model which leverages game theory alone. We also discuss graduated sanctions and how we characterize if they've emerged or not. In Section 4 we introduce our computational model and look at the simplest version (the Private Provision Model) which is absent of policy. We demonstrate agent behavior docks closely to (produces results fairly in line with) theoretical predictions. In Section 5 we introduce an agentization of a social planner (the Social Planner Model) that has flexibility in choosing a policy to solve the social dilemma. We both interpret the shape of the emergent policy under a number of conditions and discuss how the policy solutions the planner finds differs from theory. In all such cases, the social planner's policies are able to correct behavior to socially optimal levels. In Section 6, we replace the social planner with a form of democracy (the Democracy Model) where agents vote directly for representatives who have proposed their own policy solutions. We interpret the unexpected shape of the emergent policy and demonstrate that democracy can solve the social dilemma, though not socially optimally. In Section 7 we summarize and compare our results from across our subsections, characterizing what features seem pivotal in altering which types of policy solutions appear. Finally in Section 8, we conclude with the implications and propose an agenda for future work.

2 Related Literature

Our paper rests 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 as been done on the role sanctioning can play in addressing the commons problem. It's 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).

In the hundreds of case studies Ostrom analyzed of communities that successfully managed CPRs, graduated sanctions were both observed frequently and were found to be important to successful sustained management of the resource (Ostrom, 1990). Evidence of this usefulness has further solidified by many later works too, including Bardhan (1993), Ostrom (1993), Ghate and Nagendra (2005), Rubinos (2017), and van Klingeren and Buskens (2024). This is especially true when they occur with the 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 sanctioning works 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 that graduated sanctioning is more effective than strict sanctioning in the long term, and that there are specific conditions of when graduated sanctions are effective. There is also some 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. 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). 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, Unpublished Results). 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 phenomenon, which serves as the groundwork for this endeavor.

Much has also been done on modelling of coordination behavior in commons games, computational models being no exception. For example, De Geest and Miller (Unpublished Results) explore how social choice mechanisms affect the policy that emerges from agents playing a public goods game. 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). 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 number of bad actors instead of punishment for the size of the violation as we do in this study.

Our model is distinguishable from those above in a few important ways. First, instead of *population-level* selection methods with possibly random mutations to strategies (genetic algorithms, evolutionary games with replicator dynamics, etc.), we utilize *agent-level* learning. Ostrom (2014) makes the case that the evolution of rules may follow different selection processes than biological selection and that rule selection or changes to rules may be viewed in some ways as a type of 'policy experiment.'. We model this process of rule selection and learning based on 'policy experiments' to find the better performing rules. 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 Model) and citizen/agents' voting in representative democracy (the Democracy Model). 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 modelling 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 Theoretical Underpinnings

We first illustrate a game and its results when played by rational agents who make mistakes via a trembling hand (Selten, 1975), in which agents may make random mistakes from Nash Equilibrium. This theoretical result serves as the baseline against which we will compare results from our computational model with boundedly rational learning agents. We also formalize our interpretation of graduated sanctions which we will reference later when analyzing policy shapes that result from versions of our computational model.[2]

3.1 The Harvest Game

At the core of our model is the canonical N-player investment game with negative, instead of positive, externalities. This game, which henceforth we refer to as the Harvest Game,

[2] In the terminology of Ostrom's IAD framework, we are working at the "model" level but for the sake of insights at the "theory" level about how institutional features relate to one another. This is all done within Ostrom's framework in terms of agents within an action situation which, in our case, is a simulated game of a common-pool resource (Ostrom, 2014).

represents our common-pool resource and the ways in which agents can interact with it. Each time the Harvest Game is played, agents must decide how much they would like to harvest from the resource $h_i \in [0, H]$. Each unit harvested provides 1 unit of benefit to the harvesting player, but contributes to a cost which each of the agents share. In particular, the payoff to a selfish agent i as a function of their harvest choice is

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

where h_i is agent i's choice of harvest and $\overline{h}$ is the average harvest choice, i.e.

$$\overline{h} = \frac{1}{N}\sum_{i=1}^{N} h_i \tag{2}$$

Similarly, a rational, completely altruistic agent (one who cares for others' well-being as much as their own) would receive a payoff of

$$\pi_i(h_i) = \overline{h} - \alpha\overline{h} - \beta\overline{h}^2 \tag{3}$$

which is equivalent to the average selfish payoff collected by agents.

3.2 Perfectly Informed Rational Agents with Mistake Making

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 uniformly choose 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. We then take the limit as ϵ approaches 0 of these strategies to find the trembling hand equilibrium for the Harvest Game.

Solving the game, we find altruistic agents choose the socially optimal harvest level of

$$h_{SO} = \frac{1-\alpha}{2\beta} \tag{4}$$

while selfish agents choose the potentially much larger

$$h_{CE} = \frac{N-\alpha}{2\beta} \tag{5}$$

to harvest. Suppose for example there are 4 agents ($N = 4$), the maximum harvest level $H = 5$, and the parameters governing the size of the externality generated are $\alpha = 0.8$ and $\beta = 0.05$.[3] Then we'd find $h_{SO} = 2$ and $h_{CE} = 5$.

Unsurprisingly, all agents are better off under the altruistic choice (Equation 4). Since selfish agents produce sub-optimal outcomes, a natural question is whether a policy can be implemented to bring the behavior of selfish agents closer to that of altruistic agents. We investigate this in the Social Planner Model and the Democracy Model. In the theoretical framework, with rational agents who have some small chance to make an error and perfect monitoring of harvest levels, there are a set of policy solutions which provide the correct incentives to discourage rational agents from deviating from choosing non-socially optimal actions.

In both the theoretical model and in the computational Social Planner Model, a policy solution is produced by a social planner (from the top-down). In the Democracy Model, the collective actions of the agent's participation in government by voting determine the policy. In all of the aforementioned cases, a policy function f(·) is chosen which maps penalties to agents conditional on their harvest level. This would modify a selfish agent's payoff function

[3] We will use these same parameters for all computational and theoretical solutions unless otherwise noted: $N = 4$, $\alpha = 0.8$, $\beta = 0.05$, $H = 5$ and a maximum allowable fine $M = 10$.

to be

$$\pi_i(h_i, f(h_i)) = h_i - \alpha\overline{h} - \beta\overline{h}^2 - f(h_i) \tag{6}$$

while an altruistic agent would receive

$$\pi_i(h_i, f(h_i)) = \overline{h} - \alpha\overline{h} - \beta\overline{h}^2 - \overline{f} \tag{7}$$

where $\overline{f}$ is the average fine paid by agents, i.e.

$$\overline{f} = \frac{1}{N}\sum_{i=1}^{N} f(h_i) \tag{8}$$

Under the same parameters as given above, the social planner would find that the optimal policy, if utilizing fines without redistribution, is

$$f(h_i) = [0, 0, 0, 0.746875, 1.4875, 2.221875]$$

While the Nash equilibrium solution without trembling hands has an infinitely large set of possible policy solutions, the policy solution to the Harvest Game with trembling agents is simply the fine minimizing policy from that set.[4] This highlights the importance of utilizing the trembling hand solution as our point of comparison. Without the trembling hand, policies with excessive fining would all be rational policy solutions. Knowing the penalties are never realized in equilibrium, fine sizes are inconsequential in the non-trembling hand case so long as they're sufficiently large to discourage changes of behavior.

When the social planner utilizes fines with redistribution (such that there is no net loss to social welfare), any policy function which satisfies the following criteria solves the social

[4] For more details, see Appendix A.

planner's problem:

$$f(h_i) = \begin{cases} 0 & \text{if } h_i = h_{SO} \\ j \in [\max(0, A + Bh_i + Ch_i^2), \infty) & \text{otherwise} \end{cases} \tag{9}$$

where

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

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

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

Notice that this does not say much about shape of the policy itself - it only provides a lower bound on the optimal penalty levels. This means a policy vector which awards fines of size [10, 10, 0, 10, 10, 10] for $h_i = 0, 1, .., 5$ respectively solves the problem just as well as the policy function [0, 0, 0, 1, 2, 3]. Note this extends directly from the fact that while fines are painful to the agents 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's sufficiently high to deter agents from choosing something other than the socially optimal level. This result remains unchanged from the Nash equilibrium solution without trembling hands.[5]

We find that the theoretical model with (and without) mistake making is 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 (as will be made clear in the next section) and the solution for policies with redistribution indicate that 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).

[5] For more details, see Appendix A.

3.3 Defining Graduated Sanctions

Under graduated sanctions, a first, small infraction might result in a small fine, but a single large infraction or repeated offences 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 3.2. 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's 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 policy vector satisfies the criteria outlined in 9. 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.

We would expect, with limited information or bounded rationality, agents would utilize their experience from earlier 'mistakes' to formulate their long-run strategies. In such a world where mistakes/exploration choice is endogenous, the best choice of policy shape isn't so clear. While graduated policy (e.g. $f(h_i) = [0,\ 0,\ 0,\ 0.9,\ 1.8,\ 5]$) will result in smaller direct losses to social welfare due to along the learning path fines, draconian sanctions (e.g. $f(h_i)$ = [0, 0, 0, 10, 10, 10]) may correct behavior away from sub-optimal choices more quickly by sending a very clear signal in a noisy environment. If we want to contextualize the conditions that give rise to graduated sanctions, we'll need to explore a more nuanced model of the problem. We also cannot discount the role that biases and cognitive processes may play in affecting policy shapes which theory has left out. The non-trivial affect of such processes has been the focus of many researchers in psychology, biology, and behavioral economics.

Below we propose an agent-based model which allows for computational investigation of learning and in later implementations (see Section 6) provides a platform for collective decision making. The model is sufficiently flexible such that a wide array of agent-level preferences, decision-making processes, and policy-formation processes can be explored. Importantly, the flexibility with which the policy function f(.) can be chosen allows for arbitrary policy shapes to emerge, graduated, draconian, or otherwise.

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'). We intend to extend this model into dynamic punishments in future work.

4 The Private Provision Model

4.1 The Private Provision Model Description

The Private Provision Model is our first computational model. In it, we encode a simulation in which N agents play a repeated version of the Harvest Game. In each period, the agents will choose a harvest level, h_i, observe the outcome, and then collect their payoffs just as before, though now we discretize the action set so the harvest level h_i can take the value of any integer from the set {0, ... H}. We equip each of these N agents with the ability to learn from their past experiences with the goal of maximizing their individual payoff.

First, we introduce the boundedly rational learning agents we use in all three models: The Private Provision Model (this section, Section 4), the Social Planner Model (Section 5), and the Democracy Model (Section 6). For simplicity we introduce these agents once, here.

4.2 Boundedly Rational Learning Agents

Agents learn through reinforcement:[6]

- Agents store the average performance of each action.
  - Selfish agents use own utility as their measure of performance, given in equation 1 above.
  - Altruistic agents use the sum of all agent's utility payoffs as their performance metric.
- Agents have a probability p_t to explore and probability $1-p_t$ to exploit, where p_t starts large and shrinks to zero.
- When agents explore, they choose an action with a probability proportional to its average payoff. Actions which have not been chosen yet are initialized to a high level of assumed performance.
- When agents exploit, they choose the action with the highest average payoff.

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.

[6] For more details, see Appendix B.

Agents also utilize similarity as a baseline. Simply put, agents believe similar actions have somewhat 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. Similarity is a fundamental element of experiential learning. To motivate using similarity to model learning in their seminal work on case-based decision theory, Gilboa and Schmeidler (1995) quote Hume (1777): *"From causes which appear similar we expect similar effects. This is the sum of all our experimental conclusions."*

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.

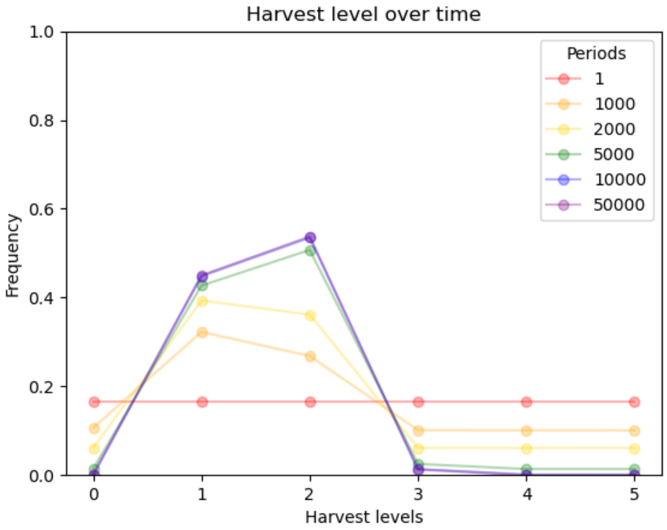
This decision-making process strikes a balance between agent sophistication and simplicity. Our agents are sophisticated enough to engage with a highly unconstrained policy function while simple enough to facilitate interpretability of decision making and direct comparison to theory.

4.3 Results

We first investigate what altruistic and selfish agents will do in our computational model when left to their own devices. Hence we run the model for populations of either fully altruistic or fully selfish agents in the absence of policy. We work with the case of $H = 5$, so there are six levels of harvest available: $0, 1, \ldots, 5$.

Altruistic agents. Recall the socially optimal choice is $h_{SO} = 2$ given the parameters detailed above. We compute the average level of attraction agents have to each of the six available harvest levels across 25 runs at different periods of the Harvest Game. We find a population of all altruistic agents chooses a harvest level of either 1 or 2 frequently. The plot below in Figure 1 illustrates the distribution of the agents' attraction to actions as it changes over time, moving in rainbow order from the initial (flat) red line to the final purple line.

Figure 1: Harvest Levels in the Private Provision Model with Altruistic Agents



Altruistic agents learning to play pro-socially.

We note (and can observe in Figure 1) 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 1, 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:

$$\overline{\Psi(X)}_{Recovered} = \frac{\overline{\Psi}(X) - \overline{\Psi}(h_{CE})}{\overline{\Psi}(h_{SO}) - \overline{\Psi}(h_{CE})} \tag{10}$$

where $\overline{\Psi}(h_{CE})$ is the average felicity generated in a round when agents all play $h_{CE} = 5$, $\overline{\Psi}(h_{SO})$ is the average felicity generated in a round when agents all play $h_{CE} = 2$, and $\overline{\Psi}(X)$ is the average felicity generated for an agent in a round 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.

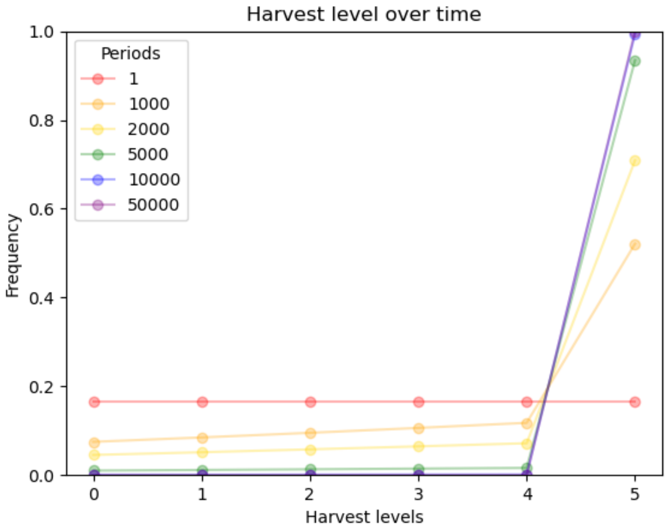
This metric normalizes the average social welfare generated in our simulation by the gap in social welfare generated in the Harvest Game when agents act altruistically vs. selfishly. If we think about the social welfare generated when agents are selfish as reality and we consider the social welfare generated 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're 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.

Selfish agents. Now recall $h_{CE} = 5$. We produce a similar plot for a population of all selfish agents below. We see agents fairly quickly converging to a harvest level of 5, demonstrating the tragedy of the commons.

Figure 2: Harvest Levels in the Private Provision Model with Selfish Agents



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

This behavior is fairly consistent with theory. Selfish agents act identically to what theory predicts in the long run. We can see that in general, agent behavior in the Private Provision Model docks fairly closely to what theory predicts.

We find that selfish learning agents also achieve a level of social welfare very similar to theoretical results, with $\overline{\Psi}(f(.), X) \approx$ -0.23 (The prediction from theory is -0.25). Again, we can use Equation 10 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.

5 The Social Planner Model

5.1 The Social Planning Model Description

In the Social Planner Model, we introduce a benevolent social planner who cares equally about the welfare of all agents in the commons and has a specific policy tool. This model modifies and extends the Private Provision Model (Section 4). Other than the introduction of this tool, the model is unmodified. Importantly, the citizen agents in the model remain the same: they are the boundedly rational learning agents described in Section 4.2.

The policy tool is a policy function f(·) which applies penalties to agents conditional on their harvest choices (just like in the theoretical model). We model the policy function f(·) as a fully flexible vector with one entry for each possible harvest level agents can choose {0, ... H}. The social planner chooses the policy function $f(h_i)$ with the goal of maximizing social welfare which is given by

$$\Psi(f(h_i)) = \Sigma_{i=1}^{N} \Sigma_{t=0}^{T} \pi_i(h_i, f(h_i)) \tag{11}$$

(Recall from Equations 6 and 7 that $f(h_i)$ enters each agent's utility as an additive penalty corresponding to their choice of h_i.)

We designed a social planner which computationally explores the fitness landscape of the policy space via a combination of hill climbing and simulated backwards induction, which we aim to summarize as follows:[7]

The social planner starts with a randomly chosen policy from the policy space. In our case, this is a vector of six numbers drawn uniformly from [0,10] (including decimals up to hundredths). Each round of the simulation, the social planner will compare their stored 'best so far' policy to close alternative policies in the policy space by running a simulation of the world under each candidate policy. The planner then observes the social welfare generated under each policy and keeps the best from that set of options.[8] This comparison to neighboring policies is repeated once for each depth of the search we allow the social planner. To be clear, the depth parameter establishes how many rounds of policy searching the social planner is engages in. Through this iterative exploration process, the social planner explores the policy space with the intention to find a high performing policy. Since this process is path-dependent, we also allow the social planner to repeat this whole process a number

[7] For more details, see Appendix C.

[8] Since any digit with two decimal places from [0,10], there are 1000 possible values for each of the six harvest levels, yielding 10^{18} possible policies

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 policy which they have found so far we denote $f^*(.)$. To ensure this exploration process wasn't ended prematurely, we look at the marginal improvements in social welfare over search-time of the policy 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 policy first. After the policy-choice stage, the agents then decide, in the next stage, how to harvest in response to the policy. 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 any policy preferences.

5.2 The Social Planning Model Results

5.2.1 The Social Planner Chooses Fines without Redistribution

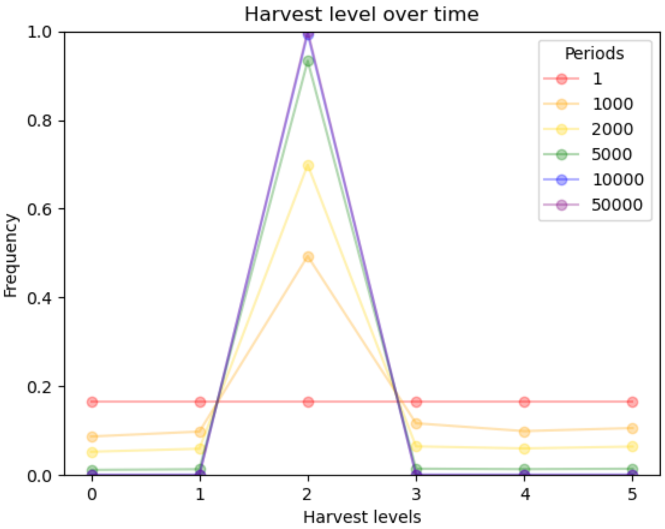
Next, we run the simulation with the benevolent social planner and selfish agents. The social planner chooses a policy in the form of a fines without redistribution, meaning the fines imposed by the policy are lost to the community—they do not directly improve social welfare. Instead, they steer behavior. 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.

In the case where agents don't use similarity in their decision making, the social planner decides to implement the following policy

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

which falls on average just 25% above the theoretical solution given in 3.2. 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 $h_{SO} = 2$.

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



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

Further, we can use a generalization of equation 10 to draw some conclusions about how well this policy solves the social planner's problem.

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

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(.). The learning agents in this case are selfish and don't use similarity in their decision making. 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 lost when agents act selfishly rather than altruistically 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.

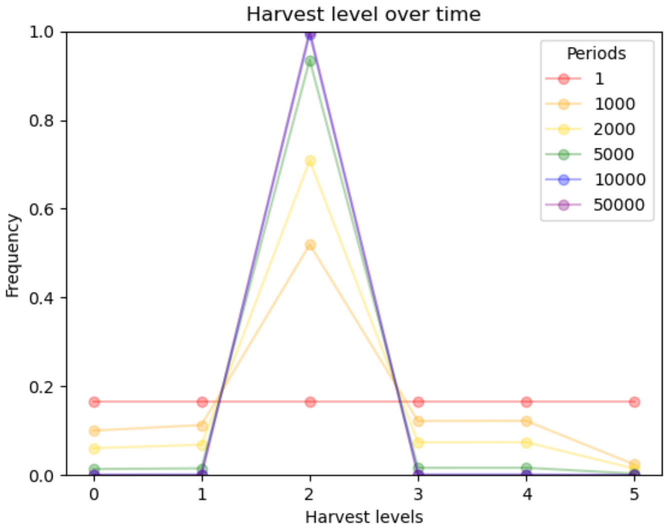
In the second case, where agents do use similarity in their decision making, the social planner finds a different-looking policy solution:

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

We can see that much of the policy looks similar to the policy solution found when agents don't utilize similarity in 5.2.1, with violations for choosing an h_i of 3 or 4 averaging to about 23% above the theoretical solution given in 3.2. 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 Model with Similarity



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

Unlike in the previous cases, it seems 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 $0 \to .99 \to .8 \to 8.3$, 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 5.2.1 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 12, we see approximately 91.5% of the social welfare gap is recovered. Interestingly, this best found policy (5.2.1) performs almost exactly as well at recovering social welfare when agents use similarity as the previously best found solution (5.2.1) 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.

5.2.2 Comparing Social Welfare Generation by the Theory Policy Recommendation

Next, we want to see how well the policy recommendation from theory (where agents have a trembling hand) 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 (5.2.1) performs more than 6 times better than the one proposed by theory.

This may be surprising to some. How can that be? Simply put, the theoretical model's agents don't 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.

5.2.3 The Social Planner Chooses Fines with Redistribution

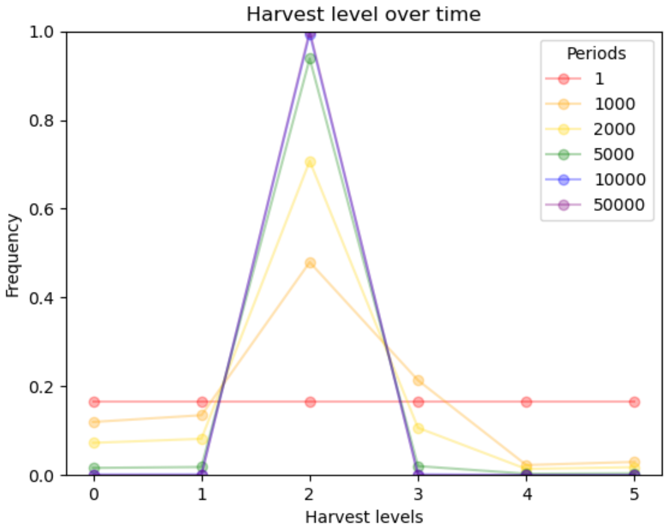
Next, we investigate a context in which the fines collected, pooled, and then redistributed, contributing directly to social welfare. Recall that, from theory, a wide array of policies could maximize social welfare equally well, as seen in Equation 9. In this context, we hypothesized that a policy vector 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 $h_i = 3$. Why? 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's 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 9. 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 Model 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.

6 The Democracy Model

6.1 The Democracy Model Description

In the Democracy Model, we investigate the effect that social choice mechanisms have on the shape and efficiency of emergent policy when we replace the social planner with an implementation of a two party representative democracy with faithful representatives. This model extends and modifies the Private Provision Model Section 4) and Social Planner Model

(Section 5); the setting is identical, and, importantly, the citizen agents in the model remain the same: they are the boundedly rational learning agents described in Section 4.2.

In this model of representative democracy, all agents participate in shaping policy by voting in elections for representative who have clearly stated policy platforms. These representatives then carry out their promised platforms without error or misrepresentation (hence "faithful."). Each election process can be summarized as follows:[9]

1. The two parties propose a point in policy space as their platform (randomly to start).

2. Agents forecast the effect the new policy will have on their own expected utility and compare it to the incumbent policy 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 proposed policy by utilizing a form of hill-climbing, informed by the number of votes they received. The winning party maintains their current platform.

In this model of democracy, parties are free to adopt different platforms (i.e. policies f(.)) with access to a completely flexible functional form. The representatives do not value the welfare of the agents directly; instead they only care about getting elected (and to a lesser extent, maximizing their vote share). This implementation is built on the idea that the pressure to offer policies which better serve voters comes from competition for votes.

This is fairly optimistic view of democracy, since, by assumption, representatives always implement they policies they campaign for, they have no personal agendas, nor do they benefit whatsoever from additional fines collected.

[9] For more details, see Appendix D.

6.2 The Democracy Model Results

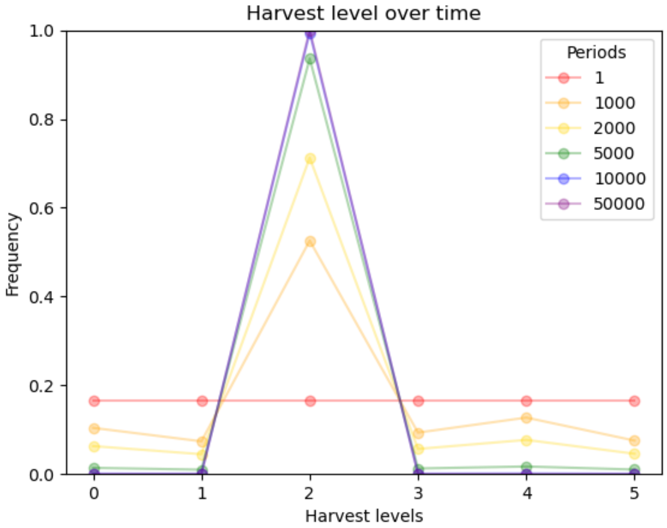
The dominant policy which emerges under our implementation of democracy is:

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

This carries some noticeably different features from policies which emerge from a top-down, benevolent social-planner in two important ways. First, the fines are not monotonically increasing in harvest levels. 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 Democracy Model



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

As we did in Section 5, we calculate how much social welfare is recovered by implementing this policy using Equation 12. 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 5.2.1), where 91.3% of the social welfare loss is recovered. The level of social welfare achieved by democracy is quite remarkable when comparing it to the top-down computational social planner which 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 88.6% 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're 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 policy.

7 Discussion

We have presented a number of variations of our computational model. In this section we take stock of what we’ve established, both summarizing what we’ve established and building upon it through comparison of our simulated experiments.

Motivating the Model

We sought to investigate whether the design principle of graduated sanctions could emerge from an agent-based model. Existing theoretical models, both with and without mistake making, are insufficient as tools to investigate elements of policy shape as discussed in Section 3, because they either do not result in graduated sanctions (which is in stark contrast to what we observe in the real world), or in the case of fine redistribution, make little claim about policy shape at all. We need a new model which encodes the important details required to study why and when we might expect to observe certain policy function shapes or features emerge as successful policy solutions in the real world. Given Ostrom's references to the potential importance of modelling the intentional exploration process of policy formation and response, we introduced a model of learning agents and policy makers. Having now seen that our model replicates theory in the simple case (with no policy) and produces distinct policy solutions under different conditions—similarity vs. not, fining with or without redistribution, social planning vs. social choice— we think our computational model has demonstrated its value as a tool for theorizing about CPRs.

Comparing the Model to Theory

We also demonstrated that in our computational model, agent behavior matched our first set of theoretical predictions in the absence of policy in Section 4.3. We also saw in both cases that a similar level of social welfare was achieved as theory predicted. The policies found in the Social Planning Model (Section 5.2) and in the Democracy Model (Section 6.2) differed from the solution found in the theoretical Harvest Game model with trembling hands. Further, we demonstrated that 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.

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 don't 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 don't 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, don't take risks with their platforms when existing ones have proven successful.

8 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. We will investigate policies with dynamic sanctioning—that is, tracking individual agents' history of violating rules and fining them accordingly. We will also investigate the role of agent heterogeneity might play in their ability to solve the commons problem, which may play a role in both how and how well the commons problem can be solved by different policy selection mechanisms. We will 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," *Theor Biol*, 2012, *299*, 152–61.

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.

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.

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.

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.

Fehr, Ernst and Simon Gächter, "Cooperation and punishment in public goods experiments," *American Economic Review*, 2000, *90* (4), 980–994.

Geest, Lawrence De and John Miller, "Using Social Choice to Solve Social Dilemmas," Unpublished Results.

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.

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.

Maja, Tavoni Alessandro Schlüter and Levin Simon, "Robustness of norm-driven cooperation in the commons," *Proc. R. Soc*, 2016, *283.*

Moor, Tine De and Annelies Tukker, "Participation versus punishment. The relationship 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.

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

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.

Rubinos, Cathy, "Commons governance for robust systems: irrigation systems study under a multi-method approach," *Doctoral dissertation - Arizona State University*, 2017.

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

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

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.

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

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

A Solving the Theoretical Harvest Game

A.1 Rational, Selfish Agents Facing No Policy:

Agents try to choose h_i to maximize their own payoff given by Equation 1, 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.

$$h_{CE} = \frac{N - \alpha}{2\beta} \tag{5}$$

A.2 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

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

A.3 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's assume the planner aims to maximize social welfare excluding the penalty to agent benefit incurred from the policy itself. Later, we'll 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

$$\pi_i(h_i, f(h_i)) = h_i - \alpha\overline{h} - \beta\overline{h}^2 - f(h_i) \tag{4}$$

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 3. 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 1 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

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

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 don't 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.

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

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 9. As we'll show in the next section, the solution set remains unchanged in the trembling hand version of the problem when fines are redistributed.

A.4 Solving the Harvest Game for Rational Agents with Trembling Hands:

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:

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

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 is a 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 5.

Similarly, an altruistic agent's problem is given by

$$\max_{\{\hat{h}_1,\ldots,\hat{h}_N | \forall \hat{h}_i, \hat{h}_i \in [0,H]\}} \Sigma_{j=1}^{N}[\hat{h}_j - \alpha\bar{\hat{h}} - \beta\bar{\hat{h}}^2] \tag{A20}$$

and once again, we find that the altruistic agent's symmetric Nash equilibrium remains unchanged (given in Equation 4) 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.

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

The policy maker's solution set (from Equation 9) 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't 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 9. Since punishment doesn't 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.

B 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. Each score in the score vector starts as some arbitrarily large number Z. 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, they agent updates the score as a running average of all past normalized payoffs observed playing the action.

Formally,

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

where

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

Intuitively, the linear transformation of utility performed to construct $\pi_i(\hat{h_i})$ 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.

3. Updating Exploration Rate:

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

$$p_{t+1} = p_t e^{-\lambda} \tag{B8}$$

C 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 25 times, each with a new initialization. The policy which performed best across all 25 simulations is the social planner's best found solution to the commons problem.

Initialization

The social planner starts with policy 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'll 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 policy 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

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

D 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

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

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 platform 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

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