Solving the Tragedy of the Commons: a model of institutional engineering for international collective action problems

Nicolas Moës

Linacre College

University of Oxford

Trinity 2017

(22,093 words)

Submitted in partial fulfilment of the requirements for the degree of the Master of Philosophy in Economics

Solving the Tragedy of the Commons:
a model of institutional engineering for international collective action problems

by Nicolas Moës, Linacre College, University of Oxford

Supervisor: *Professor Duncan Snidal, Nuffield College & Department of Politics and International Relations, University of Oxford*

Abstract

This paper develops a new approach to collective action problems and to endogenous, sustainable solutions to these issues. It shows that even in anarchic communities of rational agents with purely egoistic preferences and discounting the future, some agents have incentives to initiate the costly creation of a self-governance mechanism that solves the issue. The model demonstrates this through the analysis of a community consuming a common good. The result relies on having heterogeneous agents with satiation points, so that the tragedy of the commons unfolds over several periods and, crucially, on giving agents the ability to establish a self-governance mechanism in exchange of a transformation cost. The paper provides a new explanation to the self-governance of common-pool resources observed throughout History *that does not rely on social preferences, networks or evolution. It also makes empirical predictions consistent with some existing stylized facts derived by* Ostrom (2000). *More fundamentally, it provides a new theoretical approach to study collective action problems and the evolution of institutions.*

Contents

1. Introduction

All of the global challenges we face in the 21st century are solvable. Be it natural resource depletion, multilateral arms races, anthropogenic environmental change, wealth redistribution, trade wars, technology regulations, terrorism, vulnerability to natural disasters or underfunded research, I argue that every issue facing today's society can be resolved without technological miracles, as soon as stakeholders are willing to work together on these issues. To the extent that these issues can be analysed as public good or common goods problem under anarchy, the Economics literature provides extensive insights into the incentive schemes at play. Theory has so far focused on reasons why stakeholders *cannot* usually cooperate despite the obvious aggregate benefits. The purpose of this paper is to construct a model that would describe how, given these incentives, stakeholders facing a public good or common good problem *can* cooperate under anarchy.

To do so, I draw from the empirical research agenda that has studied self-governed communities of individuals brought together by their common use of a Common-Pool Resource (CPR) under anarchy. This line of research, led by Elinor Ostrom, has shown that the way a community manages its CPR is a critical determinant of its current welfare and is often significant for its historical development. A CPR being either a common good, a public good or a hybrid of both, traditional economic theory would predict a tragedy of the commons (Hardin, 1968) or suboptimal care-taking of the public good (Olson, 1965, Chap. I). Yet, field researchers have observed that CPR management can be sustainable, avoiding the trap of collective action problems (e.g. Ostrom, 2000).

Sustainable management in this context implies a higher level of welfare over one's lifetime, an outcome deemed desirable by the members of the community (such as sustainable development, stable peace, non-decreasing revenue streams, …). It must be distinguished from unsustainable management, whereby the CPR is depleted in such a way that none of the community members deems it desirable (resource exhaustion, conflicts, economic exodus, …). Understanding better what determines whether a community falls in the traditional economic scenario of unsustainable management or manages to overcome the incentive scheme set against its members is therefore a good starting point for solving the public good and common good problems we face today.

I develop a theoretical model that attempts to explain some findings from this empirical literature. My rationale is that if I build a theory that is consistent with these facts and common knowledge of economic behaviour, I should be more confident about predictions and policy recommendations derived from that theory about other aspects of the issues. Since a theoretical framework for solutions to collective action problem could validly apply to a vast array of issues –including many global challenges facing the international community-, the value of even a marginal improvement on current theories' predictive accuracy is non-negligible[1].

The model starts with a heterogeneous community of rational, egoistic agents consuming a common resource. Each agent has a satiation point for consumption, so that the tragedy of the commons occurs over several periods. The key innovation is to allow the agents to pay a transformation cost to create a governance mechanism (called "Arbiter") that can alter the incentives of the agents in the community. This Arbiter has its own objective function and budget constraint. I am interested in Arbiters that attempt to shift the community on a more sustainable path. The Arbiter alters incentives by allocating punishments for overconsumption.

To my limited knowledge, the establishment of such a mechanism has not been modelled in the literature, especially not in a structural fashion. Yet, it seems this exploratory version of the model is already far more consistent with empirical findings than the traditional theories predicting the tragedy of the commons and public good under-provision. Indeed, the model explains some of Ostrom (2000)'s stylized facts about successful self-organization for CPR

[1] And certainly worth the risk of a lower mark on a graduate thesis for being too exploratory.

management. Beyond that, it already provides some novel predictions about who leads a community towards self-governance and when that institutional entrepreneur does so. Specifically, the most likely community member to establish the Arbiter values the future highly compared to others and does not suffer as much from the Arbiter's punishment policy. The most likely time for establishing an Arbiter is relatively late in the process of resource depletion: at most a few periods before the full exhaustion of the common good.

Relative to the complexity of the phenomenon of self-organization, this model is too simplistic. Yet, I expect many Microeconomics theorists to consider this as too sophisticated – indeed admittedly, much remains to be done to make the analysis more straightforward. In this exploratory paper, I have tried to strip down the model from all unnecessary bells and whistles while still providing novel insights into collective action problems and a preliminary idea of policy recommendations.

My analysis proceeds as follows. In the next section, I situate my paper by reviewing the relevant literature on CPRs communities and formal institutions. I then describe the model in section 3, the core theoretical contribution of this paper. It starts with a reassessment of the tragedy of the commons, follows with a description of the Arbiter and closes with the circumstances under which a community shifts from an unsustainable to a sustainable path. For the time-constrained reader, a summary of the model is provided in section 3.6. Section 4 summarizes my early results and evaluates the model against what is known empirically about the issue. I first state some fundamental predictions radically different from traditional models of collective action by rational individuals. I then derive what characteristics make it more likely for a community to sustainably manage a CPR in section 4.2. I also briefly look at the welfare implications in section 4.3. In section 4.4, I evaluate my model's consistency with stylized facts derived from 50 years of empirical research. In section 5, I discuss a desirable but difficult extension to the model: notions of legitimacy for the Arbiter. Finally, section 6 summarizes the results and concludes.

2. Theoretical and empirical context

The literature concerned with governance of the commons under anarchy has been shaped by Nobel Prize winner Elinor Ostrom. Ever since her doctoral dissertation and case study on ground water management (Ostrom, 1965), she devoted her career to studying common-pool resources (CPRs) and their management. Alongside authors from many other academic fields, she has repeatedly provided examples of sustainable management of CPR. These resulted in widely-accepted stylized facts about factors influencing whether a CPR community is successfully self-governed. One of the key results from this empirical research is that, while some CPR communities collapse in a typical scenario of the tragedy of the commons, we often observe communities able to introduce successful self-governance mechanisms to bind community members into cooperation for the sustainable consumption of the resource (Ostrom, 2000; Gilmour et al. 2013; Acheson, 2003; Rustagi, Engel & Kosfeld, 2010; Viegas et al., 2007; and more[2])

However, since Hardin (1968)'s reasoning underlying the tragedy of the commons and Olson's model of group's cooperation for public goods (Olson, 1965, Chap. I), theoretical economists have taken a radically different perspective on collective action issues: the baseline result that they cannot be solved internally. Indeed, the almost traditional result is that rational, asocial, memoryless agents with full information cannot solve the tragedy of the commons. Additionally, because of the strategic dynamics involved, there has been a sort of methodological consensus to study this class of problems: Game Theory. Unfortunately, at that time, this field still had difficulties to generate the clear-cut predictions necessary to discriminate among models against empirical evidence, because of multiplicity of equilibria.

Naturally, until the 1990s, this widening gap between the empirical literature and theory hindered the productivity of both sides, since evidence without

[2] Such as Bardhan, 1999; Cox, 2014; Lam, 1998 and Morrow & Hull, 1996.

theory and theory without evidence generate very limited and context-sensitive policy recommendations, if any. To my knowledge, the key empirical takeaway was the list of facts derived from the detailed description of successful vs unsuccessful CPR management under anarchy, while the key theoretical takeaway was the formal analysis of incentives at play in these issues as well as significant innovations in Game Theory (exemplified more recently by Young, 2011, 2015).

Later, the advances in both research communities have led to a partial convergence. Indeed, with the development of models describing norm-building, learning and evolution, networks, bounded rationality and prosocial preferences, the theory has provided fundamental results in line with evidence: there is a non-negligible probability for a given community to solve collective action issues without government or property rights, so long as the individuals in this community are not memoryless and asocial (Castillo & Saysel, 2005). This has led empirical researchers to embrace these theoretical contributions, sometimes to the extent of recommending against using theories not relying on bounded rationality assumptions (Poteete, Janssen & Ostrom, 2010, Chap. 7)!

The rejection of these assumptions has of course led to difficulties when scholars concerned with the CPR management turned to studying formal institutions by contrast to informal ones. Indeed, informal institutions are socio-cultural norms, evolving and diffusing gradually across the community, so that the associated models of learning and network describe them accurately (Keohane & Ostrom, 1995, Introduction). These norms are difficult to change deliberately (ergo the need for learning/diffusion rather than decision-based models). More recent work has focused on allowing agents to experiment with new norms and see what works, with either a spatial diffusion via networks or a temporal one via learning (Young, 2011). From that perspective, if a community is to solve a tragedy of the commons via informal norms, it would need to either experiment a very different but beneficial norm or be originally endowed with a culture that already favours sustainability. An innovative approach to empirical research on rule-experimentation might soon provide more precise predictions to evaluate the existing theories (Rockenbach & Wolff, 2016).

To a certain extent, this paper complements this approach by making norm-experimentation endogenous and rational. Indeed, formal institutions result from more deliberate human decision-making (from deciding who is the chief to voting a law in parliament). These formal institutions might evolve into culture rather than decisions, but the crucial part is that the initial change in institution has been deliberate rather than random. A community can therefore change the current informal institutions through a deliberate and organized effort to create a new formal institution. For example, by organizing themselves, a group of advocates can become vocal enough to raise people awareness and alter the current norms of consumption prevailing in the community. Additionally, this paper should be complemented by evolutionary game theory and models of learning and network, since these accurately describe social feedback on new rules (Young, 2015), while the model presented here must still be extended to account for it.

The occurrence of organized efforts is rare because they are costly (North, 1990, pp. 86-87). Yet, their success establishes new, artificial, deliberate norms and leads to a sudden change in incentives for the entire community. Though learning from experience and network of influence can help, these two channels are unlikely to be leveraged without the presumably rational decision to experiment. This is especially true for formal institutions governing social units above the household level such as companies, villages, nations, … where interactions are so formalized that most decisions are the results of a (presumably rational) conscious reasoning.

The more recent literature on CPR management has seen the development of novel approaches to address this gap in explaining formal institutions, notably with the effort of some authors to formalize rules into a useable model (Crawford & Ostrom, 1995), with some relatively successful calibration to the available evidence. A clear enabler in that line of research has been the development of the "Attributes, Deontics, Aim, Condition, Or else" framework that allows to systematically study rules as a general concept rather than just as what they imply (Crawford & Ostrom, 1995). The more recent and more comprehensive Institutional Analysis and Development framework has also provided greater structure to the empirical research (Ostrom, 2011).

There is hope that this more formal approach to field research can be more conductive to a general theory of CPR management (Poteete, Janssen & Ostrom, 2010, Chap. 7). However, maybe because both frameworks have been developed with cross-section of communities in mind, they cannot easily accommodate the study of change in institutions over time –and notably the adoption of a new rule where there was none before. The model I present addresses this issue by formalizing the process through which a community comes to implement certain rules. Because of its rejection of the rationality assumption, it is also unclear how the current literature on CPR management can be reconciled with current economic models in fields where institutions and collective action problems matter, such as Growth Theory, Development Economics, Public Economics and Economic History. By developing a structural model reminiscent of many micro-founded theories, I hope to bridge the gap between Economics and a field that has progressively moved towards Political Theory.

Therefore, in this paper, I attempt to explain the creation of formal rules with a structural model, under the common assumptions of perfect rationality and full information. The unusual assumptions about the agents are their heterogeneity and the presence of satiation points. I recognize that the assumption of heterogeneity might lead to compatibility issues with other models in Economics. However, the assumption of homogeneity in micro-founded models of growth is often made for analytical simplicity. It is not as fundamental as the rationality and full information assumptions that are disappearing from empirical CPR models. Similarly, assuming satiation points is an analytical adjustment (supported empirically) rather than fundamental change. To my knowledge, relaxing strong monotonicity for a single good that has anyway no property rights associated with it does not contradict any fundamental microeconomic theory of the market.

As mentioned in the introduction, I evaluate my model's predictions against the available evidence. The most important findings from the empirical line of research have been neatly summarized by Ostrom (2000) into 10 stylized facts. I discuss them in section 4.4. By contrast to the procedure typical to Economics, I start from a rich set of empirical facts and derive a theoretical framework explaining these. All these findings are derived from field research, specifically case studies or small samples of local CPR communities under anarchy. While I believe formal institutions are relevant in local communities alongside informal ones, I am even more confident that formal institutions are the main rules that govern the international community and communities of social units bigger than households. Indeed, the scale and timeframe involved at these levels of interactions make learning and contagion a less accurate model of reality. Because the model presented here does not rely on these features, it therefore applies more broadly than the focus of the empirical research against which it is tested.

One might therefore be concerned about the validity of the empirical findings when applied to the international community or industries (i.e. communities of companies). However, there has been a solid bridge built between the empirical CPR literature and International Cooperation Theory in Politics (summarized in Keohane & Ostrom, 1995). There is a recognition that researchers should be on the lookout for conceptual differences, but the key outcome of this discussion is that most of the empirical principles governing anarchic communities of individuals also govern anarchic communities of nations. This has allowed the CPR literature to enrich the policy recommendations about a much broader class of problems, notably climate change (Ostrom, 2009).

The Economics and CPR management literatures have not yet developed a theory of the creation and evolution of formal institutions on its own, but this academic bridge enables an enrichment from the related International

Cooperation Theory. Introducing the International Relations and Politics literature to economists is beyond the scope of this paper, but in a spirit of interdisciplinarity, I point out to one specific contribution on the topic of rules creation and evolution. Pierson (2000) reviews the functionalist view of institutions, highlights its limits and derives several hypotheses that could help refine the theory. Some of his suggestions could help improve the model's relevance for the international community and, to a lesser extent, to local contexts. I think notably of election cycle-based time discounting and secondary objectives of the institutional entrepreneur (Pierson, 2000).

Beyond Politics, there are some promising convergences between CPR management and other literatures. To me, a convergence holding great potential that has yet to happen is with Sociology. Indeed, because of its historical focus on institutions' creation and change, the findings from Institutional Theory (in Sociology) could be productively combined with the existing stylized facts from Ostrom's research agenda (Pacheco et al., 2010). Developing a model consistent with the evidence-based conclusions from other streams of literature would reinforce its plausibility. Unfortunately, a M.Phil. thesis in Economics provides neither the time nor the space to bridge the gaps between different Social Sciences.

Standing on the shoulders of these giants, I present a model explaining why some communities establish certain formal institutions and others do not.

3. Model

The underlying goal of my research is to make accurate predictions about solutions for collective action problems, with high level of confidence. To do so, I need a model that is not only consistent with existing empirical facts but also practical to use. Indeed, if the model is consistent with empirical facts for certain aspects of the problems we focus on, we will have higher confidence in the validity of the model's predictions for other aspects of the same problems. However, the model would be of no use if it were too complex to generate testable predictions, which highlights the importance of analytical tractability. The key is therefore to properly balance accuracy of predictions and parsimony. The aim of this paper is therefore to set up such a model.

There are two related decisions I make regarding the model and guiding how it is designed. First, I will evaluate my model specifically against the Ostrom's facts on anarchic communities avoiding the tragedy of the commons, introduced above (Ostrom, 2000). As explained in section 2, these have been accepted across the field as stylized facts and represent the only systematic collection of observed design features of successful self-governance mechanisms. The model is therefore designed to reflect these features. Second, I assume every agent in the model is perfectly rational and has full information. These are fundamental assumptions that greatly simplify the calculations. From my own attempts at integrating incomplete information in the model, I found out that the math involved was too much so, especially given the timeframe allowed for this thesis.

Given the unusual structure and evolution of the model, I build it incrementally. I first start by characterizing the object of the collective action – here, a common good, but resolving a public good problem would also be feasible with this model[3]. I then define the agents, their characteristics and objectives. From there, I can describe the baseline scenario: a tragedy of the commons. I then explain what happens when the agents can self-organize and to create a new agent that I call Arbiter. The Arbiter comes with a certain amount of power to affect the other agents' consumption decisions. In my experience, models usually do not allow for such a "meta-structural" transformation, but that is the core innovation of this paper, so I detail it extensively. I first describe the Arbiter's objective, budget constraint and its consumption targets agenda. This allows me to derive the decisions it makes to nudge the agents towards a more

[3] As discussed in Bowles (2004), pp. 130-131, the only analytical difference between both types of good is that the public good's level of common stock enters the utility of agents directly and public good "consumption" is negative and represents a contribution.

sustainable consumption path. I then explicitly derive the conditions under which the Arbiter is created. I finally put all these results together in a summary of the model.

To my very limited knowledge, this is the first attempt at endogenising agent-driven change in the structural settings of the model. Yet, I believe modelling such a change is the only way to properly describe change in formal institutions and solutions to collective action problems.

3.1 Common resource

There is a good S assumed to exist ex ante, produced by nature. I assume nobody has property rights over this good or at least nobody can enforce his property rights over it. Consumption of the good is therefore non-excludable, to the extent that you have physical access to the good. Consumption is however rival, meaning that what is consumed by someone cannot be consumed by someone else. S is therefore a common good. S could represent fisheries stocks in international waters over which nations have no authority, the stock of clean atmosphere or, from a local perspective, a forest in an area where the local government cannot enforce the rule of law. I even argue that S could represent the level of goodwill between nations, implying that consumption is any action (statement, policy, event, …) by a nation that reduces the overall cooperative feeling and trust in the international community. Since the model does not require to be specific about what the good is, I carry on with good S and let the reader free to choose what real concepts it represents best.

In terms of notation, I denote the stock of the common good S at the beginning of period t by S_t, with $S_0 > 0$ and $S_t \geq 0$. The time dimension t extends from period 0 to the final period T, which can be infinity. I assume that S_t naturally replenishes (disappears) at the rate r which is positive (negative) and constant over time. r cannot be smaller than -1. This rate can also be zero or infinitely small in human timescales (as in the case of minerals and fossil fuels), but many classical examples of common goods have non-zero growth rates – at least for some time.

Indeed, when left untouched, fisheries stocks and forests grow or regress until their ecological carrying capacities are reached and the atmosphere converges towards its baseline "natural" composition everywhere. I must nevertheless assume r is constant over time for analytical tractability, despite being aware that this growth rate varies in reality. Finally, the core of the problem arises from the fact that the good S is consumed every period by n agents (described below). The agents' aggregate consumption at period t is denoted by $\sum_j^n s_{t,j}$.

For simplicity, I assume the agents consume before the stock's periodical growth, so that only the remaining stock grows[4].

Given the replenishment and the consumption, the law of motion for S_t (the way the stock of good S evolves over time) is given by:

$$\Delta S_t = -\sum_j^n s_{t,j} + r \, * \left(S_t - \sum_j^n s_{t,j}\right) \tag{1}$$

With $S_0 > 0, S_t \geq 0$ for $t = 1, 2, \ldots, T$ and $\Delta S_t = S_{t+1} - S_t$. Or, similarly:

$$S_{t+1} = (1 + r) * \left(S_t - \sum_j^n s_{t,j}\right) \tag{2}$$

Tomorrow's stock is therefore an increasing function of today's starting stock and of the growth rate, and a decreasing function of today's aggregate consumption. This is a typical common good.

[4] One could also imagine the growth to occur before the agents consume the good (giving a law of motion $r\, S_t - \sum_j^n s_{t,j}$). However, this simple permutation of events makes the derivations more complex.

3.2 Community members

I assume that there are n different agents who have access to good S and who cannot be excluded from consuming it, with $1 < n < +\infty$. Together, they form what I refer to as a community. These agents can be individuals, households, villages, companies, associations, parties, nations, groups of nations, … any social unit making decisions regarding a common good. For example, these can be fishermen exploiting a lake's fisheries, governments exploiting goodwill among nations for domestic purposes, oil companies exploiting a common oil field or researchers abusing trust of society. To avoid extreme cases where a single agent's decision determines the outcome, I assume the community is made up of agents of the same class (i.e. all agents are individuals, all agents are companies or all agents are nations). It should be noted that the model allows for so-called "multiclass" situations where e.g. individual fishermen interact with companies, corporations interact with nations or researchers interact with associations. Within the common class, I nevertheless assume there is heterogeneity of preferences and extraction technologies.

Indeed, the n agents are indexed by $i = 1, \dots, n$. During period t, agent i extracts $s_{t,i}$ from S_t and consumes this amount straight away. To avoid first-mover advantages, I assume all the agents consume simultaneously. Because consumption of S is rival, the physical limit on $s_{t,i}$ is:

$$s_{t,i} \leq S_t - \sum_{j=1, j\neq i}^{n} s_{t,j} \tag{3}$$

Agent i has preferences over the quantity of good $s_{t,i}$ that she consumes. Within each period, these preferences are described by the utility function $u_i(s_{t,i})$, where $u_i(.)$ is real-valued, twice differentiable on $(0, \check{s}_i)$, strictly increasing and strictly concave over its domain $[0, \check{s}_i]$. These conditions are necessary for being able to find an optimal consumption level per period with the optimization methods used.

Specifically, the condition of being real-valued ensures that we can convert the quantities consumed in reality into intuitive (and mathematically convenient) quantities of utility. It implies that the agent can evaluate different levels of consumption and state her preferences about these. The condition of twice differentiability ensures that we can find the marginal change in utility from a marginal change in consumption, which is necessary for ensuring that a local optimum is found in the dynamic problem. It implies that the agent can rank all consumption levels and that a marginal change in consumption will lead at most to a marginal change in his welfare (no jumps in utility). The condition of being strictly increasing is sufficient for the problem of overconsumption to occur. It also ensures that each level of utility is associated to at most one level of consumption. It implies that the agent always prefers more of the good rather than less and that he is never indifferent between 2 different levels of consumption.

Finally, the condition of strict concavity is necessary to ensure that at most one level of consumption is selected as optimal within each period. It is crucial for obtaining a trade-off between consuming now and consuming later, which makes dynamic optimization problems relevant in the first place. It implies that the more the agent consumes within one period, the less enjoyment she derives from any additional consumption of the good within that period (diminishing marginal utility from consumption within each period). These are common assumptions in economic models so I make them for analytical tractability, even though experimental evidence on some of these conditions is sometimes mixed.

The domain of $u_i(.)$ is limited below to $0 \leq s_{t,i}$ to avoid trading cases whereby agents who highly value the good "buy it away" from agents who produce (i.e. negatively consume) it in exchange of rewards from the Arbiter, rewards that come from punishments of the buyers (the Arbiter is described in greater details below). While potentially valuable, a theory of market design based on this Arbiter model is beyond the scope of this thesis. The implication of nonnegative values of $s_{t,i}$ is that while agents can extract from S or abstain, they cannot produce S.

More importantly, the domain of $u_i(.)$ is also limited above to $s_{t,i} \leq \check{s}_i$. Given that the function is strictly increasing, $\check{s}_i$ should be understood as the within-period satiation point of agent i with respect to good S. It is an admittedly artificial way to integrate a satiation point in the utility function, but it is the simplest way that allows to analytically describe the Tragedy of the Commons. The real counterpart of $\check{s}_i$ is the consumption quantity beyond which the individual or nation finds that consuming more of the good is effectively impossible, e.g. for technological reasons. Given the complexity already built in the model, I prefer to avoid assuming non-monotonic functions, which would have been necessary to endogenise the satiation point.

Note that the satiation point is different for every agent, but is constant over time. It is quite natural to think that different agents have different needs and extraction technologies. Different satiation points imply different $u_i(.)$ across agents, but there can be other sources of difference in the form of $u_i(.)$. Regardless, the form of $u_i(.)$ does not depend on time – preferences are stationary across periods[5].

[5] The stationarity of preferences is a necessary assumption for the dynamic programming methods used to solve the model. It must be criticised since empirically it is more attractive to consume the common goods in some periods rather than others (e.g. elections period, recessions, drought year, certain seasons, …).

Furthermore, agent i considers today's utility gains or losses as more important than tomorrow's. She therefore discounts utility in the next period by applying the discount factor β_i, which is strictly positive, strictly smaller than 1 and constant over time. $\beta_i < 1$ is necessary to ensure that the agent prefers consuming now, so the Tragedy of the Commons can occur in the model.

Given these within-period features, I assume that the agent has a lifetime utility from the common good spanning from period 0 to period T given by: [6]

$$U_i(S) = \sum_{t=0}^{T} [\beta_i^t * u_i(s_{t,i})] \qquad (4)$$

Expressing the lifetime utility as the sum of discounted "within-period" utility is not a trivial decision. It assumes that the lifetime utility is additively separable over time. This means that consumption of the good during period t does not affect directly utility at any other period than t. Specifically, the agent does not derive utility in period $t+1$ (or $t-1$) from remembering (or expecting) consumption at period t. This assumption of temporal additive separability is crucial to solve this class of dynamic problems analytically.

[6] with T allowed to be infinity – which is again contested empirically.

With the elements developed here and in the preceding section, I can already write down formally the problem faced by the agents. Agent i's objective is to maximize her lifetime utility by choosing $s_{t,i}$, subject to the constraints from the common good and the assumptions explained above. Formally:

$$\max_{0 \leq s_{t,i} \leq \check{s}_i} U_i(S) = \sum\nolimits_{t=0}^{T} \left[\beta_i^t * u_i(s_{t,i})\right]$$

$$subject\ to$$

$$S_{t+1} = (1+r) * \left(S_t - \sum\nolimits_j^n s_{t,j}\right) ; \quad S_0 > 0 ; \quad S_t \geq 0 \tag{5}$$

$$s_{t,i} \leq S_t - \sum_{j=1, j\neq i}^{n} s_{t,j} \quad and \quad 0 \leq s_{t,i} \leq \check{s}_i$$

Note how the stock S_{t+1} is determined not simply by i's consumption, but rather by the sum of all the other agents' consumption. This makes the consumption decision inherently strategic –i's decision is affected by the others' decisions and vice versa. I explore this in the next section.

3.3 The tragedy of the commons

I now explain what happens when agents are consuming good S. I first develop the case in which the agents are naïve to derive key benchmarks for my model and to clearly emphasize where the collective action problem comes from. I then build on this solution by allowing strategic behaviour to radically transform the results.

3.3.1 A community of naïve agents:

I assume for now that all the agents are "naïve" i.e. they cannot behave strategically. They do not consider how their decisions affect (and are affected by) others' decisions. Problem (5) that each naïve agent faces is therefore a so-called "cake-eating problem" with a growing cake S_t and exogenous time-varying downward jumps (due to others' consumption of the cake). A solution to this problem is a rule of behaviour or a policy: a function for agent i that maps the observed value of the stock S_t to an optimal value of consumption $\hat{s}_{t,i}$ for every t. Given the rules of behaviour of all agents, I can then describe how S evolves over time. Assuming all the agents are naïve and given the conditions on $u_i(s_{t,i})$, this rule of behavior is of the form:

$$\hat{s}_{t,i} = \begin{cases} 0 & if \;\; f_i[\beta_i, r, (S_t - \sum_{j=1, j\neq i}^{n} \hat{s}_{t,j})] < 0 \\ f_i[\beta_i, r, (S_t - \sum_{j=1, j\neq i}^{n} \hat{s}_{t,j})] & if \; 0 \leq f_i[\beta_i, r, (S_t - \sum_{j=1, j\neq i}^{n} \hat{s}_{t,j})] \leq \check{s}_i \\ \check{s}_i & if \;\; f_i[\beta_i, r, (S_t - \sum_{j=1, j\neq i}^{n} \hat{s}_{t,j})] > \check{s}_i \end{cases} \quad (6)$$

To ensure that the naïve agent must restrain herself from consuming –and so to ensure that there is a welfare reduction from strategic interactions later in the model- I focus on the cases where $\hat{s}_{t,i} \leq \check{s}_i$. This assumption also ensures the solution is consistent with the domain of $u_i(.)$, so that I can use dynamic programming to solve the problem[7].

By solving the dynamic optimization problem of each agent, I obtain a system of n equations with n unknowns. A numerical solution therefore exist, but if the within-period utility functions of the different agents are identical up to a linear transformation, I can untangle the system analytically and obtain $\hat{s}_{t,i}$ as a function of exogenous parameters $(\beta_i, \beta_{j\neq i}, r, S_t)$, where $\beta_{j\neq i}$ is the vector of discount factors of the other agents in the community:

[7] The trick is to use $(S_t - \sum_{j=1,j\neq i}^{n} \hat{s}_{t,j})$ as state variable. I would like to thank Godfrey Keller for his advices on the use of dynamic programming for this class of problems

$$\hat{s}_{t,i} = \hat{s}_{t,i}(\beta_i, \beta_{j\neq i}, r, S_t) \tag{7}$$

The assumptions made in the previous sections –notably the naivety assumption- ensure that (7) corresponds to usual cake-eating solutions. *Ceteris paribus* the direct effect of a higher β_i makes the agent values the future consumption more, so that current consumption will be relatively lower. *Ceteris paribus,* a lower average $\beta_{j\neq i}$ makes the others' current consumption bigger, so that there is less available for agent i. In a naïve community where others' consumption is considered exogenous, there is an "offsetting", whereby an agent that is more future-oriented relative to the others' average restrains its own consumption to try to bring the aggregate consumption down and have more to consume later.

The effect of r is ambiguous. On one hand, since *ceteris paribus* a greater r ensures a greater future level of S, so that consumption in the future is higher, implying that it is optimal to consume more currently to smooth this effect. On the other hand, a greater r also means that greater restraint today is relatively more rewarding, via a higher level of S in the future – there would therefore be an incentive to reduce consumption. The direction of the net effect of an increase in r is therefore dependent on the empirical circumstances. And finally, of course, *ceteris paribus* a higher S_t leads to a net increase in individual current consumption levels.

The transition rule of the common good in this naïve community will therefore be:

$$S_{t+1} = (1+r) * \left(S_t - \sum_{i}^{n} \hat{s}_{t,i}(\beta_i, \beta_{j\neq i}, r, S_t)\right) \tag{8}$$

(7) and (8) provide us with an optimality benchmark. Naivety achieves the same result as if property rights were optimally attributed to agents at the beginning of the game: if each agent received a certain fraction of S to manage on her own[8] at $t = 0$, she would optimally smooth consumption over time and obtain the same lifetime utility as if she was in a naïve community with no property rights. Naivety as defined here is therefore sufficient to prevent the tragedy of the commons to occur.

3.3.2 A community of strategic agents:

I have shown in the previous section that the naïve agent's decision depends on the other agents' decisions (cf. equation (6)). This illustrates the possibility of strategic interactions in this community. These strategic interactions are the reason why most of the literature on commons goods problems focus on game

[8] This specific fraction would of course depend on each agent's discount factor relative to others'.

theoretical models. However, a key novelty in this paper is to carry on with a structural approach. Indeed, since I have integrated a satiation point for the agents, there is a determinate structural process underlying the tragedy of the commons.

To emphasize the individual incentives and the process along the way, I first describe the situation when there is only one strategic agent in the community and the others still behave naively. Starting from (6) and given that the restraint condition holds, a strategic agent in a naïve group has incentives to consume more than what the naïve agent's policy would command. Indeed, since (6) holds for all agents, she knows that the rest of the community will simultaneously restrain themselves even more to avoid diverging too much from the naïve aggregate consumption path. This is the offsetting discussed in the previous section. Specifically, if i is the strategic agent, we have:

$$\hat{s}_{t,i}^{non-naive} = \hat{s}_{t,i} + x_{t,i} \tag{9}$$

And, integrating this into (6) for the other agents:

$$\hat{s}_{t,j} = f_j\left[\beta_j, r, \left(S_t - x_{t,i} - \sum_{k=1,k\neq j}^{n} \hat{s}_{t,k}\right)\right] \quad for\ all\ j \neq i \tag{10}$$

The offsetting takes place in (10): the extra consumption of agent i directly reduces the other agents' consumption. Effectively, by consuming a marginal amount $x_{t,i} = \iota_{t,i}$ extra, the total stock decreases only by $\iota_{t,i} - \sum_{j\neq i}^{n} \frac{\partial f_j}{\partial(S_t - \sum_{k=1,k\neq j}^{n} \hat{s}_{t,k})}$. The strategic agent enjoys the full marginal utility gain from extra consumption, without the full marginal lifetime utility loss. The other agents therefore provide a utility subsidy to the strategic agent. The offset leads the strategic agent to optimally chose a positive $x_{t,i}$.

When all agents are strategic though, the offsetting disappears while the extra consumption remains. Let's start with all agents being strategic but currently following their naïve policy. Each agent has an incentive to defect and pick an arbitrary positive $x_{t,j}$, since given the others' policy this decision increases their lifetime utility. Given that any agent deviates, no strategic agent has incentive to offset by restraining her own consumption. All agents therefore "defect" and overconsume. Reiterating this reasoning, all agents would regret having picked any $x_{t,j}$ that does not make $\hat{s}_{t,j}^{non-naive} = \check{s}_j$, the upper limit on $\hat{s}_{t,j}$. The Nash equilibrium has therefore all agents consuming their satiating consumption:

$$\hat{s}_{t,j}^{non-naive} = \check{s}_j \qquad (11)$$

Given the restraint condition, this consumption policy exhausts the common good at a faster rate than in the naïve situation and, S disappears earlier than in a naïve community (unless r is big enough relative to $\sum_j^n \check{s}_j$ to ensure S is *never* exhausted). This is the tragedy of the commons. Given the assumption of full information, strategic agents will realize this "unescapable" destiny in period 0 and will therefore not even attempt to restrain themselves at any period (restraint is never a Nash equilibrium since deviations are rewarding). The associated transition rule for the good is therefore:

$$S_{t+1} = (1 + r) * \left(S_t - \sum\nolimits_j^n \check{s}_j \right) \tag{12}$$

(11) and (12) the baseline benchmark. Given the restraint condition, (11) is greater than (7), so that the transition rule (12) implies a faster decrease than (8).

Depending on the size of S_0 and r, the agents enjoy a certain number of periods with satiating consumption at the beginning of the game. At a certain period, denoted t^*, the sum of satiating consumption levels is greater than the remaining stock from the previous period. This is the last period at which agents will extract an amount from the good, since afterwards the stock is exhausted. Note that if the size of the stock is low enough relative to the satiation points, $t^* = 0$. In the cases of international fisheries and fossil fuels extraction, the tragedy of the commons spans over decades. In this model, the reason why international fisheries are not exhausted under anarchy is not due to restraint, but to a within period satiation point. Fishing more than this satiation point is simply unprofitable given current demand and technologies. It therefore makes sense to build this satiation point in the model.

There is no single way to distribute the remaining stock at t^* because agents consume simultaneously. I could assume that they split the good evenly among themselves or, if we have a specific form of within-period utility function I could assume that those who derive the most utility from the good "fight the hardest" to get more of it in the last period. This would imply that the agents distribute the good among themselves up to the point where the marginal utility of each is equal to the others'.

When does the stock runs out? t^* is the first period at which the remaining stock is nonnegative but smaller than the aggregate sum of satiating consumption levels. Using the transition rule (2), this implies the following condition (cf. Appendix 1.1):

$$\min t^* : 0 \leq S_{t^*} = (1+r)^{t^*} S_0 - \frac{(1+r)^{t^*+1} - (1+r)}{r} \sum_i^n \check{s}_i < \sum_i^n \check{s}_i \qquad (13)$$

This condition results in the following values for t^*, depending on the growth rate, where for notational clarity $X = \sum_i^n \check{s}_i$ is the aggregate satiating consumption (cf. Appendix 1.1):

$$t^*(r, S_0, \sum\nolimits_j^n \check{s}_j) = roundup(\varepsilon^*) \tag{14}$$

With the values for ε^* given in table 1 where, for notational clarity, $X = \sum_i^n \check{s}_i$. A greater aggregate satiating consumption level (greater X), a smaller original stock (smaller S_0) or a smaller growth rate all imply an earlier depletion of resources.

Table 1: Values determining t as a function of r*

$\boldsymbol{r}$	→	$\boldsymbol{\varepsilon^*}$
$r = -1$	→	0
$-1 < r < 0$	→	$\frac{\ln(X) - \ln(X - r(S_0 - X))}{\ln(1 + r)}$
$r = 0$	→	$\frac{S_0 - X}{X}$
$0 < r < \frac{X}{(S_0 - X)}$	→	$\frac{\ln(X) - \ln(X - r(S_0 - X))}{\ln(1 + r)}$
$\frac{X}{S_0 - X} \leq r$	→	$+\infty$

It should be noted that, in reality, there might be some self-restraint occurring under anarchy. The cognitive, social and informational aspects of the situation might push agents to consume less than satiating amounts. Indeed, if decision-makers have limited strategic reasoning, they might not realize that their actions indirectly affect others and thus consume as if everybody did the same. An agent might have social preferences that lead her to internalize the costs of their consumption to the community. Trust and the ability to communicate with other decision-makers have also been shown experimentally to increase cooperation in collective action problems (Ledyard, 1995; Chaudhuri, 2011). Finally, the uncertainty about future situations, other agents' preferences and intentions will make the agent hesitant to initiate a tragedy of the commons, depending on the agent's risk preferences. My model assumes away these behaviours with the perfect rationality and full information assumption as well as the egoistic form of the utility function.

I now have a workable model of the tragedy of the commons. I can focus on developing a workable solution to the issue. The interesting cases for this model is when $t^* < T$ (unsustainability) and when there is no central authority able to force the agents onto a sustainable path (anarchy). When both anarchy and natural unsustainability hold, the community members end up locked in a suboptimal outcome because of their strategic incentives.

To escape this outcome, game theorists often suggest "communication" as a remedy that could allow them to reach a sustainable outcome, e.g. by making a coordinated effort to permanently reduce $\sum_i^n \check{s}_i$ (Chaudhuri, 2011). However, the CPR management literature and Ostrom's empirical work in particular offer some insights into what "communication" be and, contrary to experimental settings, it is much more than communication that is needed in most real contexts: the agents must set up formal institutions and rules.

3.4 Arbiter

The key innovation in this paper is to model the endogenous creation of an additional agent that I call "Arbiter". This Arbiter represents the formal institution(s) created by anarchic communities worldwide to tackle specific collective action problems. In reality, since it arises under anarchy, its authority can only be derived from the consensus of the community members. However, as a first introduction to the self-governance mechanism, I assume that the Arbiter's power is exogenous. For now, the Arbiter is therefore characterized by a given level of power and by a community onto which it exerts its authority. Examples of Arbiters include many different forms of organization across time and communities, from the agreed-upon monitor of water consumption in a local irrigation system to the administrative systems making up the UN. While many of these mechanisms deal with multiple collective actions issues simultaneously, I focus on a single issue here.

I explain how the Arbiter works and what its objectives and constraints are. I then detail what consumption agenda it tries to enforce: first, what aggregate consumption it allows, then how it distributes this quota among the agents. I then describe how it achieves a certain level of compliance among agents regarding this agenda.

3.4.1 Arbiter's objective and budget:

The Arbiter is characterized by an objective, a budget and the decisions it makes given these, which I call "sanctions". In this section I describe the objective and the budget. I must first make the distinction between "agenda" and "enforcement". The Arbiter decides on a consumption *agenda* for the community, which is a list of targets consumption level for each agent at each period. Its objectives in agenda-setting are exogenous in this version of the model, but could be made endogenous, as I discuss in section 3.4.2. The Arbiter is also characterized by an *enforcement* objective, which I describe below. This enforcement objective is the guideline for how to sanction agents that do not respect the agenda. As I discuss, enforcement and agenda-setting can be source of politicking among agents designing the Arbiter and preference aggregation or bargaining models could be used to make this politicking endogenous.

The enforcement objective of the Arbiter is for the agents to consume exactly the amount that each is allowed to by the Arbiter's agenda. I use a standard loss function, which penalizes deviations from a given target of the Arbiter. If $s_{t,i}^{*}$ is the Arbiter's optimal consumption to be imposed on agent i and $s_{t,i}$ is the observed consumption by the agent, the loss function of the Arbiter is:

$$\mathrm{L}(s_{t,i}) = -\sum\nolimits_{i}^{n} (s_{t,i}^{*} - s_{t,i})^2 \tag{15}$$

Observe that the squaring penalizes both under- and overconsumption. However, I have shown previously that there are only overconsuming agents in the community.

The squaring implies that, given a binding budget constraint, the Arbiter attributes sanctions in priority to agents generating relatively bigger deviations from the agenda's path and will tolerate small deviations. This loss function is given exogenously. I assume this is the only loss function "design" available to the community. Of course, in reality the agents negotiate among themselves to determine what the Arbiter's objective is. While it could be interesting to endogenise the choice of the objective as a direct result of the coalition's preferences and bargaining, I do not model this process for the sake of parsimony.

There are two implicit assumptions underlying the features of this loss function. First, the Arbiter is impartial: it cares equally about deviations from any agent, since for each term of the sum, the loss from a certain deviation away from the target is the same across agents. This is a strong assumption in the case of international commons, since in reality the Arbiter might be created by some nations to constrain other geopolitically-threatening nations. Impartiality is appealing analytically, since it provides symmetric sanctions to all deviating agents. However, a fair objective does not imply a fair Arbiter. Indeed, I discuss below how the Arbiter determines the target consumption levels for each agent and I show that this allocation of targets can be "unfair". I simply assume away partiality of enforcement with this symmetric loss function.[9]

Second, the loss function is such that the Arbiter cares about all n agents, regardless of whether some of these agents are in favour of having an Arbiter in the first place. By this assumption, the Arbiter has therefore incentives to correct the consumption behaviour of all the agents. To be consistent, I therefore assume that the Arbiter has authority over all agents in the community. It is possible to model cases where the monitor has more or less difficulty sanctioning certain agents by attributing an agent-specific factor between 0 and 1 to scale down the sanctions of some agents e.g. those that do

[9] To develop asymmetric sanctions, one could scale each agent's deviation in L(.) by an agent-specific coefficient.

not recognize the Arbiter's authority. This does not add much intuition to the analysis, so I do not integrate this feature in the model.

The Arbiter is also characterized by a budget constraint. The budget here should not be considered as a financial budget but rather as a "power" or "influence" budget, that can be spent onto dissuading agents from consuming beyond what is sustainable. In this model, the power of the Arbiter is therefore to affect the within-period utility level of any agent. In reality, this might be the power to publicly shame a community member who abuses the local irrigation system or to instore economic sanctions on a nation's government officials. Later, I discuss an extension of the model where the Arbiter's power is dependent on how many agents support the Arbiter.

For now, I just assume that the Arbiter's power is exogenous and given by P, where $0 < P$. The Arbiter therefore expends its power by imposing dissuasive sanctions to agents. The sanctions incentivize the agents to restrain their consumption the way the Arbiter wants, but the sanctions are costly in terms of power for the Arbiter to impose. I assume that for each unit of "dissuaded consumption" (i.e. sanction) $p_{t,i}$ inflicted on agent i, the Arbiter incurs a loss of μ in power for that period. μ is an exogenous parameter representing the Arbiter's efficiency at dissuading consumption.

The Arbiter sanctions agents every period and is limited in doing so by the following budget constraint:

$$\mu \sum_{i}^{n} p_{t,i} \leq P \tag{16}$$

Note again that by assuming the existence of a $p_{t,i}$ for each agent among the n agents of the community, I effectively assume that the Arbiter has influence over all agents. (16) is a simplistic non-dynamic budget that allows me to keep the model tractable.

The objective function and the budget are the core characteristics of the Arbiter. I now describe its decisions regarding what target consumption levels to pick for all agents and, then, what policy it uses in terms of sanctions to maximizes its objective given the budget constraint.

3.4.2 Agenda-setting:

The Arbiter is useful in solving the collective action issue insofar as it has different incentives from those of the agents. Indeed, as an Arbiter, it does not face the strategic environment experienced by the community members. The Arbiter can therefore take a decision regarding the aggregate consumption of the community (denoted s_t^*) without worrying about externalities and then allocate the consumption targets (denoted $s_{t,i}^*$ for all i) to the agents in the community. While it can therefore avert the tragedy of the commons, whether this solution to the collective action problem leads to an improvement in aggregate lifetime welfare depends on what consumption agenda the Arbiter sets.

The target aggregate consumption s_t^* could be given exogenously, or the Arbiter could be given exogenous temporal preferences over the common good, or -for the most enthusiastic- the Arbiter's preferences themselves could be endogenised in a bargaining game among the community members. The important change to highlight is that even if the Arbiter has preferences, its optimization problem is not corrupted by strategic incentives. It follows the same decision-process as an agent in a "community" of $n = 1$. If these preferences satisfy the conditions imposed on the agents' preferences, its optimization results in a smoothing of consumption over time until T.

Therefore, to remain tractable and yet extract relevant insights from the model already, I assume that the Arbiter is exogenously given preferences. The Arbiter design available to the community is therefore assumed to have a discount factor $0 < b < 1$. It can be endogenised in several ways: b could be an aggregation of the coalition's time preferences β. The agents might have agreed on b after a vote, after multilateral bargaining, randomly or via any decided mechanism. The within-period preferences attributed to the Arbiter is $u_A(s_t^*)$ and satisfy the same condition, except it has domain $[0, \sum_i^n \check{s}_i]$. I denote by x the period at which the Arbiter is created, so that it is effective at $x+1$. Finally, the Arbiter has an additively separable lifetime utility, from $x+1$ to T.

The Arbiter then decides on the consumption path $s_t^*(S)$ to maximize his attributed utility function. This consumption path determines an optimal aggregate consumption for each period, which can then be allocated following the distribution rule chosen by the coalition (discussed below). Following the reasoning for the naïve agent and assuming the Arbiter's restraint condition holds, I obtain:

$$s_t^* = f_A[b, r, S_t] \tag{17}$$

for $t > x$. Having an expression for the aggregate consumption each period, the Arbiter's second decision is about how to allocate it.

The Arbiter decides how it distributes the targets $s_{t,i}^*$ among the community members given the aggregate consumption level s_t^* decided for each period. One can think of many ways to formalize this allocation and, again, the way ultimately selected in reality depends on the members negotiating the rules shaping the Arbiter. One could imagine that the allocation is simplistic, so that each agent receives $\frac{s_t^*}{n}$ of the target aggregate consumption, regardless of individual characteristics. Alternatively, the Arbiter could be assumed to be a welfare-maximizing utilitarian and take the within-period preferences of each agent into account. In that case, it would allocate a share of s_t^* to each agent such that the marginal utility for each agent is equal across all agents, that is $u'_i(s_{t,i}^*) = u'_j(s_{t,j}^*)$ for all i and j, subject to $\sum_i^n s_{t,i}^* = s_t^*$. Interestingly, we could also imagine a corrupt Arbiter, where some agents with certain ethnic, religious, familial, or socioeconomic characteristics are allowed very generous consumption targets and others must suffer most of the burden from restraining aggregate consumption. However, in that case the model becomes a model of government's social contract and kleptocracies rather than a model of self-governance for collective action.

In any case, the target must be physically achievable by the agent. Given the domain limits on each agent's utility function, the boundaries on the targets are $0 \leq s_{t,i}^* \leq \check{s}_i$. Since it cannot set a target for an agent beyond her satiating consumption, this implies the Arbiter always aims at a consumption path that is not less sustainable than the situation under anarchy.

3.4.3 Arbiter's sanctions:

I now turn to the way the Arbiter can enforce this path. As introduced when setting up the budget constraint, every period the Arbiter has n decisions $p_{t,i}$ to incentivize the agents, with $p_{t,i} \geq 0$. I assume here that these are punishments. The model works equally well with rewards, but empirically, punishments are more common. These punishments directly affect the agents' within-period utility functions, as if converted into common good-equivalent units. Agents now maximize $u_i(s_{t,i} - p_{t,i})$. In this world with complete information, I effectively assume the Arbiter is fully credible about the punishment, given its budget. I also assume that agents marginally prefer not being punished given the same utility level. This means that even if $s'_{t,i} = \hat{s}_{t,i} - p_{t,i}$, an agent systematically prefers consuming $s_{t,i} = s'_{t,i}$ over consuming $s_{t,i} = \hat{s}_{t,i}$ and being punished $p_{t,i}$. This avoids the uninteresting requirement of having to threaten marginally more than the deviation from target.

I further stretch the limits of simultaneity by assuming that, within each period, the n agents simultaneously decide on their optimal strategic consumption $\hat{s}_{t,i}$ (which, is constant at $\check{s}_i$ in this model), are then dissuaded by the Arbiter via threat of punishment $p_{t,i}$ and therefore instantaneously re-adjust their actual consumption to $s'_{t,i} = \hat{s}_{t,i} - p_{t,i}$ (given their marginal preference for not being punished). The current punishment mechanism therefore does not especially reflect a physical expense by the Arbiter, but rather a credible threat of punishment. The budget constraint must be respected for the Arbiter's threats to be fully credible in this world of perfect information. This is the simplest structure I found to enforce targets, although again, imagination is apparently the only limit, given the great diversity of institutional engineering observed throughout history.

Nevertheless, the simple punishment mechanism engineered here can accommodate several types of incentivizing actions observed in reality – and could model several of these simultaneously. A joint statement about climate change by the IPCC can be represented by an Arbiter inducing a small, temporary decrease in utility onto all nations proportional to the deviation from the consumption agenda, representing the small reputation cost to polluting governments arising from temporarily-increased awareness of the issue. By adapting the agent-specific weights of its loss function, the Arbiter might find it optimal to heavily sanction a small subset of the community, as is the case with economic sanctions against certain nations. To model a blanket ban, the loss function might be such that the Arbiter's target is $s^*_{t,i} = 0$ for all i, t, with a lump-sum penalty reflecting a fine. Similarly, a fee scheme or tax scheme can be implemented by making the sanctions dependent on the level of past consumption. Many other institutional arrangements could be interesting to compare within this framework.

Each of these real-life arrangements for sanctions can be modelled by imposing a structure on the matrix of dimensions $n * (T - x)$ containing all $p_{t,i}$. However, in this paper the only condition I impose on it is $p_{t,i} \geq 0$ for all i, a fairly general case. Additionally, because of the domain condition on $u_i(.)$, the optimal $p_{t,i}$ must also satisfy $0 \leq \hat{s}_{t,i} - p_{t,i} \leq \check{s}_i$. The upper bound is guaranteed to hold with $p_{t,i} \geq 0$. The lower bound implies $p_{t,i} \leq \hat{s}_{t,i}$. However, given the loss function and the nonnegative punishments, this lower bound is always satisfied for $0 \leq s^*_{t,i} \leq \check{s}_i$.

Because of the sanction and the intention to consume more than what the Arbiter allows, the agent gets $u_i\left(\hat{s}_{t,i} - p_{t,i}\right) = u_i\left(s'_{t,i}\right)$ and therefore consumes $s'_{t,i} = \hat{s}_{t,i} - p_{t,i} \leq \hat{s}_{t,i}$ since, as mentioned earlier, $u_i\left(s'_{t,i}\right)$ is monotonically increasing and for identical levels of utility the agent marginally prefers not being punished. The Arbiter can therefore moderate an agent's consumption by punishing her.

Putting all the elements described earlier together, the Arbiter faces the following problem for every period:

$$\max_{p_{t,i}\forall i\in[1,\dots,n]} \mathrm{L}\left(s'_{t,i}\right) = -\sum\nolimits_{i}^{n}\left(s^{*}_{t,i} - s'_{t,i}\right)^{2}$$

$$subject\ to$$

$$\mu\sum\nolimits_{i}^{n} p_{t,i} \leq P \tag{18}$$

$$s'_{t,i} = \hat{s}_{t,i} - p_{t,i} \qquad for\ all\ i$$

$$0 \leq p_{t,i} \leq \hat{s}_{t,i} \qquad for\ all\ i$$

If the budget constraint of the Arbiter does not bind, the solution is straightforward: it sets the punishments exactly equal to the deviation from the agenda. That is, if $\mu\sum_{i}^{n}(\hat{s}_{t,i} - s^{*}_{t,i}) \leq P$, then $p^{*}_{t,i} = \hat{s}_{t,i} - s^{*}_{t,i}$ (cf. Appendix 1.2).

The solution when the budget constraint is binding is more involved. Indeed, the Arbiter allocates its limited dissuasive power to minimize its loss function. It distributes its sanctions so that the marginal reduction in deviation from targets due to a marginal increase in $p^{*}_{t,i}$ are equalized across all deviating agents. If $\mu\sum_{i}^{n}(\hat{s}_{t,i} - s^{*}_{t,i}) > P$ with C "cheaters" (indexed by c) and n-C "non-cheaters" (indexed by n-c) in the community, the solution is then (cf. Appendix 1.2):

$$p_{t,c}^{*}=\frac{1}{C}\left[\frac{P}{\mu}+(C-1)\left(\hat{s}_{t,c}-s_{t,c}^{*}\right)-\sum\nolimits_{j=1,j\neq c}^{C}\left(\hat{s}_{t,j}-s_{t,j}^{*}\right)\right]$$

$$for\ all\ c\ \in[1,2,\dots,C]$$

(19)

$$and$$

$$p_{t,n-c}=0\ \ for\ all\ n-c\in[C+1,\dots,n]$$

We see that if the Arbiter is budget constrained, an agent who intends to deviate from her target more than the average deviation from other cheaters will suffer a greater punishment relative to others. There is therefore some form of gradation of sanctions, proportional to how severe the deviation is. Naturally, given the Arbiter's finite budget, the greater the number of cheaters, the smaller the individual punishment. If there wasn't a hard satiation point, this would lead to incentives to "gang up" against the Arbiter, though ganging up is a collective action problem in itself. I do not explore it here.

Except at period t^{**} where the stock is too small for consumption after the Arbiter's sanctions, in the situation explored in this paper, all n agents intend to consume satiating consumption. Indeed, consuming below satiation decreases their lifetime utility and is therefore not a Nash equilibrium (as explained in section 3.3). With c = n, (19) collapses to one case. The aggregate consumption path followed by the community after the Arbiter's dissuasion is therefore:

$$\sum_{i}^{n}\left(\hat{s}_{t,i}-\frac{1}{n}\left[\frac{P}{\mu}+(n-1)\left(\hat{s}_{t,i}-s_{t,i}^{*}\right)-\sum_{j=1,j\neq i}^{n}\left(\hat{s}_{t,j}-s_{t,j}^{*}\right)\right]\right) \qquad (20)$$

in periods where the budget binds, and

$$\sum_{i}^{n} s_{t,i}^{*} \qquad (21)$$

in periods where the budget does not bind. (20) and (21) summarize the situation under the Arbiter's regime. Given that $0 \leq s_{t,i}^{*} \leq \check{s}_{i}$, $P/\mu > 0$ and $p_{t,i} \geq 0$, aggregate consumption under the Arbiter is smaller than $\sum_{i}^{n} \check{s}_{i}$, which is what the agents consume when they are not self-governed (cf. (11)).

It should be noted that when deriving consumption target endogenously, one must make a clear distinction between agenda-setting and enforcement. The Arbiter cannot develop consumption targets based on whether it is able to enforce them, otherwise the Arbiter could effectively choose its loss function to make sure it is de facto maximized. The optimization process of the agenda must be separate from the enforcement decisions, not unlike the separation of power between the legislative and executive branches of governments.

The Arbiter does not fully enforce its targets when its budget is strictly smaller than the aggregate deviation from targets in the community, scaled by its efficiency. Nevertheless, the Arbiter still slows down the tragedy of the commons by decreasing the agents' consumption towards the target consumption (if the chosen consumption path is indeed different from permanent satiating consumption for all, of course). If the consumption targets are derived endogenously, a binding budget implies that the stock of the common good decreases faster than what the Arbiter presumed when setting up the agenda, especially since agenda-setting is assumed separate from enforcement concerns. Depending on how the targets have been endogenised, they might not be function of any level S but rather a deterministic amount as function of t. As time progresses, the Arbiter's plan might therefore be inconsistent with the community's reality and would therefore require regular updating. Since it is not necessary for the targets to be endogenous to have a workable model, I do not implement an updating mechanism. I just assume that once set, the targets remain, even if the state of the world has diverged from the Arbiter's plan.

Therefore, the period at which S runs out for a community with a budget-constrained Arbiter is $t^{**} \leq T$, given the Arbiter's potential incapacity to fully enforce its targets. Again, I can assume a sharing rule for when the good runs out at time t^{**}. It is possible to compute t^{**} from the Arbiter's agenda, punishment, budget and efficiency, as well as the agents' satiating consumption. More concretely, t^{**} is the first period at which actual consumption (given dissuasion) is greater than that period's initial stock. It is more complicated to express t^{**} as a function of these parameters.

Indeed, when the target consumption levels are time-varying and without other assumptions about how they are generated, there seems to be no concise expression for t^{**}. One way out is to assume that the Arbiter sets targets constant over time. In that case, I can express t^{**} as a function of the other parameters. The analysis is similar to the one I did for t^{*}, explained in Appendix 1.1. Since the targets are now time-invariant by assumption, the punishment is $p_{t,i} = p_i$. I can compute the value of t^{**} as follows:

$$t^{**}(r, S_x, \sum_{j}^{n}(\check{s}_j - p_j)) = roundup(\varepsilon^{**}) \tag{22}$$

With the values for ε^{**} are reported in table 2 where, for notational clarity, $Y = \sum_i^n(\check{s}_i - p_i)$ is the aggregate consumption when the Arbiter exists and where $S_x = S_0(1+r)^x - \sum_j^n \check{s}_j \sum_{z=1}^{x}(1+r)^z$. The assumption of time invariant targets is restrictive, but fortunately the model does not rely on it. Indeed, a value of t^{**} can be found given the Arbiter's targets and budget, the satiating consumption of the agents, the stock of the good and its growth rate.

*Table 2: Values determining t** as a function of r*

$\boldsymbol{r}$	→	$\boldsymbol{\varepsilon}^{**}$
$r = -1$	→	0
$-1 < r < 0$	→	$\frac{\ln(Y) - \ln\,(Y - r(S_x - Y))}{\ln\,(1 + r)}$
$r = 0$	→	$\frac{S_x - Y}{Y}$
$0 < r < \frac{Y}{(S_x - Y)}$	→	$\frac{\ln(Y) - \ln\,(Y - r(S_x - Y))}{\ln\,(1 + r)}$
$\frac{Y}{S_x - Y} \leq r$	→	$+\infty$

3.5 Conditions for creating the Arbiter

As briefly mentioned above, the novelty of this model is to give agents the possibility to establish a self-governance mechanism called Arbiter. In reality, agents negotiating the Arbiter's creation face a 2-step procedure: first, they must agree that something should be done to stop the tragedy of the commons – i.e. they must agree that some self-governance mechanism is beneficial to solve the collective action problem. Second, they must design the mechanism and therefore agree on what alternative outcome to select (e.g. should the Arbiter be fair or corrupt? How should it value future utility? …). This is reminiscent of Libecap's (1995) distinction between agreeing that there are gains from reforming the institutional settings and agreeing about the distributional implications of the proposed reforms. For simplicity, I restrain the set of possible institutional designs available to the community to the one with a "fair" loss function and the punishment mechanism described above. Given this assumption, Libecap's second step is exogenous to the model.

The single-design assumption is restrictive - probably the most restrictive assumption of the entire model. Indeed, throughout history, society has been governed by a variety of formal institutions for different collective action problems. A significant drawback from the single-design assumption is that I cannot study competition among self-governance mechanisms, which could have made testable predictions about the persistence of institutions. For example, if the existing Arbiter somehow disadvantages certain agents, these disadvantaged agents would have strong incentives to engineer their own competing institution[10]. This can only be shown if I extend the set of possible designs to more than 1. The model being in its preliminary phase, I prefer to carry on with the single-design assumption to carefully detail the creation of a given Arbiter rather than extending the model to allow multiple Arbiters.

[10] A notorious example of such a manoeuver is the creation of the Asian Infrastructure Investment Bank that resulted from China's frustration with how the Bretton Woods institutions distributed the development projects across nations (Editorial, *The Guardian,* 2014). Seeing project-based economic development and its finite pool of projects as a common good, one nation being allocated a project prevents another one from obtaining the project (and thus from expanding its influence in the target developing area).

An uncontroversial assumption in the neo-institutional literature is that transforming the institutional landscape is costly (North, 1990, pp 86-87). This explains the formal institutions' persistence despite their perceived inefficiencies in contemporary environments. I therefore assume an Arbiter is created when an agent i pays the necessary transformation cost e_i. Under anarchy, this cost reflects the amount of time and effort necessary to reach out to other agents and explain the idea. Indeed, whether one is a fisherman worried about the local lake's fisheries or a delegation leader at the UN worried about humanity's food supply, agents are busy and surely have other productive use of their time than trying to gather fellow agents to discuss collective action issues and to explain the idea of an Arbiter. Also, the social and political pressure might be such that it is difficult to break conformity: it is a psychological cost to advocate for change when nobody in the community has done so before. To ensure proper discussions, it is necessary to have the proper context: gathering people in the same place and ensuring everybody can communicate. This also involves some logistical costs. With the uncertainty characterizing sovereign nation states' reality, this also represents the political cost of surrendering some authority over domestic policy-making to an outside entity, which is geopolitically weakening.

To reflect these costs, when agent i creates the Arbiter at period x, she incurs a 1-period transformation or institutional engineering cost e_i to set up the institution, with $0 \leq e_i \leq \check{s}_i$. One could assume a different cost structure. For example, in some contexts it is more natural to think of this transformation requiring a first initiator to pay e_i, and following supporters to pay $e_j/2$. The model being already quite complex as it is, I assume the transformation costs are paid by the first agent only. e_i is converted in units of common good, so that it enters the within-period utility function the same way as $\check{s}_i$ does, but with a negative sign.

I now explain the agent's decision of whether to pay the institutional engineering cost. Each fully informed agent i faces 2 specified states of the world: one in which there is no Arbiter and one in which she creates an Arbiter in period x at cost e_i, so that it becomes effective in period x+1. Given full information about these 2 potential scenarios, each agent asks herself every period whether to create the Arbiter. The agent therefore compares whether lifetime utility with the Arbiter created (and the cost e_i paid) in period x is greater than lifetime utility without the Arbiter. If the following condition holds for agent i, she has ex-ante incentives to create the Arbiter at period x:

$$\begin{gathered}\sum_{t=0}^{t^*-1}[\beta_i^t u_i(\check{s}_i)] + \beta_i^{t^*} u_i(\hat{s}_{t^*,i}) + \sum_{t=t^*+1}^{T}[\beta_i^t u_i(0)] \\ \leq \\ \sum_{t=0}^{x-1}[\beta_i^t u_i(\check{s}_i)] + \beta_i^x u_i(\check{s}_i - e_i) + \sum_{t=x+1}^{t^{**}-1}[\beta_i^t u_i(\check{s}_i - p_{t,i})] \\ + \beta_i^{t^{**}} u_i(\hat{s}_{t^{**},i} - p_{t^{**},i}) + \sum_{t=t^{**}+1}^{T}[\beta_i^t u_i(0)]\end{gathered} \tag{23}$$

I can re-express this condition more intuitively as a minimum level of utility required when paying e_i in period x:

$$\begin{gathered}\beta_i^x u_i(\check{s}_i - e_i) \geq \\ \sum_{t=x}^{t^*-1} \beta_i^t u_i(\check{s}_i) + \beta_i^{t^*} u_i(\hat{s}_{t^*,i}) + \sum_{t=t^*+1}^{t^{**}} \beta_i^t u_i(0) - \sum_{t=x+1}^{t^{**}-1} \beta_i^t u_i(\check{s}_i - p_{t,i}) \\ - \beta_i^{t^{**}} u_i(\hat{s}_{t^{**},i} - p_{t^{**},i})\end{gathered} \tag{24}$$

Since $u_i(.)$ is a continuous increasing function, its inverse is also a continuous increasing function. I can re-express this condition as follows:

$$\begin{gathered}e_i \leq \\ \check{s}_i - u_i^{-1}\left(\frac{1}{\beta_i^x}\left[\sum_{t=x}^{t^*-1} \beta_i^t u_i(\check{s}_i) + \beta_i^{t^*} u_i(\hat{s}_{t^*,i}) + \sum_{t=t^*+1}^{t^{**}} \beta_i^t u_i(0) \right.\right. \\ \left.\left. - \sum_{t=x+1}^{t^{**}-1} \beta_i^t u_i(\check{s}_i - p_{t,i}) - \beta_i^{t^{**}} u_i(\hat{s}_{t^{**},i} - p_{t^{**},i})\right]\right)\end{gathered} \tag{25}$$

I call this the Condition for agent i to Create the Arbiter (CCA$_i$). Note that CCA never holds at $x = t^*$, since then the good is exhausted and creating an Arbiter just generates a useless cost. Given the restricted range of the within-period utility function, its inverse function is defined only when its argument belongs to $[u_i(0), u_i(\check{s}_i)]$. This condition is only necessary to express the CCA$_i$ in consumption terms rather than utility terms and does not invalidate (24).

Naturally, every agent knows that other agents also have incentives to create the Arbiter at period x or in a period soon after x, such that the agent might have strategic incentives to let another agent pay the cost. With full information, each agent can evaluate how her utility would be in all the scenarios where another agent creates the Arbiter in her stead. She therefore knows exactly whether her lifetime utility would be higher by free-riding. This is a form of asymmetric volunteer dilemma. In the words of Douglass North, changing formal institutions requires "substantial resources or at the very least overcoming the free-rider problem" (North, 1990, p. 87), which can prove difficult.

Given the heterogeneity in the community it is possible that for at least one of the agents the scenario where she derives the most utility is the scenario where she creates the Arbiter herself. This would mean it is her dominant strategy to volunteer to create the Arbiter[11]. In that case, the community is guaranteed to eventually create an Arbiter when her CCA_i is satisfied. Diversity in real communities is so significant that it could well explain occurrences of institutional entrepreneurs overcoming the volunteer's dilemma throughout history. The heterogeneity observed renders the volunteer's dilemma less relevant. Specifically, denoting by x_i the time periods at which CCA_i holds and

[11] If several agents have this dominant strategy, the one who can implement this strategy the earliest in time is the one creating the Arbiter, since she is the first one to act on her decision.

by t_i^{**} the period at which S is exhausted in the presence of the Arbiter created by agent i, the dilemma does not occur when the following condition holds:

$$\begin{gathered}
\exists\, \mathrm{i} \in \{1, \dots, \mathrm{m}\} : \ \forall\, \mathrm{j} \in \{1, \dots, \mathrm{m}\}, \mathrm{j} \ \neq \mathrm{i} \\
\sum_{t=0}^{x_i-1} [\beta_i^t u_i(\check{s}_i)] + \beta_i^{x_i} u_i(\check{s}_i - e_i) + \sum_{t=x_i+1}^{t_i^{**}-1} \left[\beta_i^t u_i\left(\check{s}_i - p_{t,i}\right)\right] \\
+ \ \beta_i^{t_i^{**}} u_i\left(\hat{s}_{t_i^{**},i} - p_{t_i^{**},i}\right) + \sum_{t=t_i^{**}+1}^{T} [\beta_i^t u_i(0)] \\
> \\
\sum_{t=0}^{x_j-1} [\beta_i^t u_i(\check{s}_i)] + \beta_i^{x_j} u_i(\check{s}_i) + \sum_{t=x_j+1}^{t_j^{**}-1} [\beta_i^t u_i\left(\check{s}_i - p_{t,i}\right)] \\
+ \ \beta_i^{t_j^{**}} u_i\left(\hat{s}_{t_j^{**},i} - p_{t_j^{**},i}\right) + \sum_{t=t_j^{**}+1}^{T} [\beta_i^t u_i(0)]
\end{gathered} \tag{26}$$

where m refers to the number of agents for whom CCA holds at some point in the game (note that this is simply a new indexing of the subset of agents for whom CCA holds). I can re-express it to obtain a condition on the minimum utility level at time $\mathrm{x_i}$ when paying the transformation cost:

$$\begin{gathered}
\exists\, \mathrm{i} \in \{1, \dots, \mathrm{m}\} : \ \forall\, \mathrm{j} \in \{1, \dots, \mathrm{m}\}, \mathrm{j} \ \neq \mathrm{i} \\
\beta_i^{x_i} u_i(\check{s}_i - e_i) > \\
\sum_{t=0}^{x_j-1} [\beta_i^t u_i(\check{s}_i)] - \sum_{t=0}^{x_i-1} [\beta_i^t u_i(\check{s}_i)] + \beta_i^{x_j} u_i(\check{s}_i) \\
+ \sum_{t=x_j+1}^{t_j^{**}-1} \left[\beta_i^t u_i\left(\check{s}_i - p_{t,i}\right)\right] - \sum_{t=x_i+1}^{t_i^{**}-1} \left[\beta_i^t u_i\left(\check{s}_i - p_{t,i}\right)\right] \\
+ \beta_i^{t_j^{**}} u_i\left(\hat{s}_{t_j^{**},i} - p_{t_j^{**},i}\right) - \ \beta_i^{t_i^{**}} u_i\left(\hat{s}_{t_i^{**},i} - p_{t_i^{**},i}\right) \\
+ \sum_{t=t_j^{**}+1}^{T} [\beta_i^t u_i(0)] - \sum_{t=t_i^{**}+1}^{T} [\beta_i^t u_i(0)]
\end{gathered} \tag{27}$$

This condition ensures that no volunteer dilemma arises, so I denote it NVD. NVD is more likely to hold when $e_{i}{}_{\mathrm{i}}$ is smaller and the earlier CCA$_{\mathrm{i}}$ holds relative to the other agents' CCA, consistent with Diekmann (1993)'s empirical finding that the one who loses the least from volunteering relative to others is most likely to volunteer. The time at which CCA$_{\mathrm{i}}$ holds ultimately depends on the agents' preferences, so that the differences in timing depend on how heterogeneous preferences are.

If this condition is not satisfied for a certain community, the agents are unwilling to create an Arbiter because their maximizing behaviour induces them to wait for somebody else to make the effort. The typical Nash equilibrium would therefore be mixed, with a small probability of volunteering. A community of agents who are too similar won't have a clear volunteer willing to bell the cat. As mentioned above, this an asymmetric volunteer's dilemma. The strategic analysis of volunteer's dilemmas has led to inroads, but these cannot yet be used in a structural context with full information (see Diekmann (1985) for his seminal investigation and Myatt & Wallace (2008) for an evolutionary perspective[12]). In particular, analytical results for asymmetric volunteer's dilemma are contradicted by experiments (Diekmann, 1993).

[12] As well as Weesie (1994) and Feldhaus & Stauf (2016) for experimental results and considerations.

This deadlock can be escaped in two ways. First, one could use a Schelling point by assuming that e.g. the agent who satisfies her CCA earliest in time or the agent with the greatest difference between the lifetime utility derived in both situations creates the Arbiter. The latter Schelling point seems supported in experimental settings, at least among European university students (Diekmann, 1993; Przepiorka & Diekmann, 2013). However, it would be foolish to claim this finding applies to nation states or even companies. Otherwise, given the full information, one could assume that the opportunity cost from *not* free-riding is internalized in e_i. This effectively assumes agents have figured it out, even though it cannot be described explicitly. While appealing for the model's tractability, this trick makes e_i difficult to operationalize for empirical applications.

The presence of a volunteer's dilemma when NVD does not hold weakens the model's predictive power. The indeterminacy arising from the mixed strategy Nash equilibrium and the small sample currently available to researchers pre-empt falsifiability. One could assume a Schelling point (based on evidence discussed above) or limit the use of the model to situations where NVD holds. NVD is sufficient to ensure that CCA is strictly necessary for an Arbiter's creation. However, NVD is not necessary to observe the creation of an Arbiter, since there is a mixed strategy Nash equilibrium for the dilemma.

My intuition is that the model therefore applies better to collective actions among nations than to collective actions among individuals or organizations within a nation. Indeed, given the existing diplomatic relations and transnational forums, most of the fixed costs necessary for nations to create an Arbiter are already sunk (notably because they have been established to create formal institutions for previous collective action problems), even though the nations still evolve in an anarchic community. At sub-national levels, there generally isn't this pre-existing network – at least not in communities that are anarchic, since I expect anarchy to disappear after successfully developing sustainable self-governance mechanisms.

Moreover, the CPR literature has mostly been concerned with resource extraction for a certain industry (villages of fishermen, irrigation systems for crop production, international fisheries, clean atmosphere…). In these cases, I believe that individuals within a local industry have similar technologies of extraction, education and similar resource use. However, nations in the international community differ greatly in terms of technology, domestic institutions, and resource use, even in a globalized world, so that preferences vis-à-vis of environmental change or international fisheries are more heterogeneous. We have seen this heterogeneity might be strong enough to ensure NVD holds.

Finally, the international community is a community only for the past 3 centuries while local communities result from much longer evolutions. Local communities divide, disperse and merge over time, but nations are "stuck together" in the international community and cannot alter their geographic locations on a given planet. I therefore expect preferences and extraction technologies to have converged and harmonized further in local communities than in the international community, since the former have evolved to be communities today, while the latter is here regardless of convergence.

Because of higher heterogeneity and lower institutional engineering costs relative to potential gains, I therefore suspect the volunteer's dilemma to be less of an issue in the international community than in the local communities. This is merely an opinion based on these arguments, of course.

3.6 Summary of the model

I have a common resource S described by the following equations:

$$S_{t+1} = (1+r) * \left(S_t - \sum_j^n s_{t,j}\right) \tag{1*}$$

$$S_0 > 0 \; and \; S_t \geq 0 \tag{2*}$$

With S_t the remaining stock of the resource at time t and $-1 \leq r < +\infty$ being the natural growth rate of S. I have a community of n rational agents who maximize lifetime utility over periods 0 to T given by:

$$U_i(S) = \sum_{t=0}^{T} [\beta_i^t * u_i(s_{t,i})] \;\; for \; all \; i \in \{1, \dots, \mathrm{n}\} \tag{3*}$$

With $0 < \beta_i < 1$ being the agent's discount factor, $s_{t,i}$ being her consumption from the common resource S, and $u_i(.)$ being her within-period utility function, which is real-valued, twice differentiable, strictly increasing and strictly concave over its domain $[0, \check{s}_i]$. This implies a satiation point $\check{s}_i$ for each agent. Given the agent's perfect rationality and the strategic context, her optimal consumption is her satiating consumption $\check{s}_i$ - this is the tragedy of the commons.

Given the agents' aggregate satiating consumption, S might eventually run out. The period at which S runs out is denoted t^*. t^* is a function of the growth rate, the initial level of S and the aggregate satiating consumption:

$$t^*(r, S_0, \sum_j^n \check{s}_j) = roundup(\varepsilon^*) \tag{4*}$$

ε^* is given in table 1 where $\sum_j^n \check{s}_j = X$ for notational clarity.

Table 1: Values determining t as a function of r*

$\boldsymbol{r}$	→	$\boldsymbol{\varepsilon^*}$
$r = -1$	→	0
$-1 < r < 0$	→	$\frac{\ln(X) - \ln(X - r(S_0 - X))}{\ln(1 + r)}$
$r = 0$	→	$\frac{S_0 - X}{X}$
$0 < r < \frac{X}{(S_0 - X)}$	→	$\frac{\ln(X) - \ln(X - r(S_0 - X))}{\ln(1 + r)}$
$\frac{X}{S_0 - X} \leq r$	→	$+\infty$

The agents can however create an Arbiter at time $x + 1$ if they are willing to pay the necessary transformation cost e_i at time x. The agent-specific e_i is given exogenously and constant with $0 \leq e_i \leq \check{s}_i$. The Arbiter is characterized by a budget constraint (i.e. power constraint) given by:

$$\mu \sum_i^n p_{t,i} \leq P \tag{5*}$$

where $p_{t,i}$ is the size of the dissuasive sanction that the Arbiter inflicts upon agent i at time t, with $p_{t,i} \geq 0$. μ is the coefficient of efficiency of the Arbiter's sanctions. P is the exogenous dissuasive power of the Arbiter. The Arbiter has authority over the entire community of n agents. It spends its power budget to minimize its loss function, the sum of squared deviations between what its target for each agent and what each agent actually consumes:

$$\mathrm{L}(s_{t,i}) = -\sum_{i}^{n} (s_{t,i}^{*} - s'_{t,i})^2 \qquad (6^*)$$

where $s_{t,i}^{*}$ is the Arbiter's target for agent i at time t, with $0 \leq s_{t,i}^{*} \leq \check{s}_i$. $s'_{t,i}$ is what the agent consumes after being incentivized by the Arbiter. The sanction directly enters the agents' within-period utility functions, so

$$s'_{t,i} = \check{s}_i - p_{t,i} \qquad (7^*)$$

where $\check{s}_i$ is what agent i would consume at time t if there were no Arbiter. The Arbiter's targets $s_{t,i}^{*}$ can be endogenous if we are willing to give the Arbiter's preferences.

As derived in section 3.4.3, the Arbiter distributes its sanctions optimally given its budget, specifically:

$$p_{t,i}^{*} = \frac{1}{n}\left[\frac{P}{\mu} + (n-1)\left(\check{s}_i - s_{t,i}^{*}\right) - \sum\nolimits_{j=1,j\neq i}^{n}\left(\check{s}_j - s_{t,j}^{*}\right)\right] \tag{8*}$$

when its budget binds and

$$p_{t,i}^{*} = \check{s}_i - s_{t,i}^{*} \tag{9*}$$

when its budget does not bind.

The time at which the good S runs out in the presence of an Arbiter is denoted by t^{**}, which can be recovered from the other parameters of the model (cf. section 3.4.3)

Given the cost e_i, agent i has incentives to create the Arbiter when the Condition for agent i to Create the Arbiter (CCA$_i$) holds:

$$\begin{aligned} e_i &\leq \\ \check{s}_i - u_i^{-1}&\left(\frac{1}{\beta_i^x}\left[\sum_{t=x}^{t^*-1}\beta_i^t u_i(\check{s}_i) + \beta_i^{t^*} u_i\left(\hat{s}_{t^*,i}\right) + \sum_{t=t^*+1}^{t^{**}}\beta_i^t u_i(0)\right.\right. \\ &\left.\left. - \sum_{t=x+1}^{t^{**}-1}\beta_i^t u_i\left(\check{s}_i - p_{t,i}\right) - \beta_i^{t^{**}} u_i\left(\hat{s}_{t^{**},i} - p_{t^{**},i}\right)\right]\right) \end{aligned} \tag{10*}$$

Given the potential volunteer’s dilemma, one can impose an additional condition sufficient to guarantee that an Arbiter is created whenever CCA holds for at least one agent. This condition is called “No Volunteer Dilemma” (NVD) and implies that there exists at least an agent for whom the utility maximizing decision regarding the Arbiter’s creation involves creating the Arbiter herself, given what the others’ agent utility maximizing timing of the creation is. This condition translates into the requirement that the community is heterogeneous enough so that at least one agent has a dominant strategy to volunteer. This can be formalized as follows:

$$\begin{gathered}
\exists\, \mathrm{i} \in \{1, \dots, \mathrm{m}\} : \forall\, \mathrm{j} \in \{1, \dots, \mathrm{m}\}, \mathrm{j} \neq \mathrm{i} \\
\beta_i^{x_i} u_i(\check{s}_i - e_i) > \\
\sum_{t=0}^{x_j-1} [\beta_i^t u_i(\check{s}_i)] - \sum_{t=0}^{x_i-1} [\beta_i^t u_i(\check{s}_i)] + \beta_i^{x_j} u_i(\check{s}_i) \\
+ \sum_{t=x_j+1}^{t_j^{**}-1} [\beta_i^t u_i(\check{s}_i - p_{t,i})] - \sum_{t=x_i+1}^{t_i^{**}-1} [\beta_i^t u_i(\check{s}_i - p_{t,i})] \\
+ \beta_i^{t_j^{**}} u_i\left(\hat{s}_{t_j^{**},i} - p_{t_j^{**},i}\right) - \beta_i^{t_i^{**}} u_i(\hat{s}_{t_i^{**},i} - p_{t_i^{**},i}) \\
+ \sum_{t=t_j^{**}+1}^{T} [\beta_i^t u_i(0)] - \sum_{t=t_i^{**}+1}^{T} [\beta_i^t u_i(0)]
\end{gathered} \qquad (11^*)$$

where x_j refers to any period where CCA$_{\mathrm{j}}$ holds and m is the number of agents for whom CCA holds at some point during the game.

4. Implications and evaluation

This model provides an analytically tractable description of anarchic communities facing a collective action problem about a common good. This new approach has several implications. First, I can derive some general results about solutions to the tragedy of the commons based on the model. Within each community, the outcomes of interest are whether the community creates an Arbiter and the impact the Arbiter has on social welfare. I discuss these in turn. I then turn to evaluating the model against Ostrom's facts.

4.1 General implications

Although the model is complex and has two restrictive assumptions, there are some key results that apply broadly. First, the key insight is that *heterogeneous communities of rational, fully-informed, asocial, memoryless agents with satiation points can resolve the tragedy of the commons by establishing costly formal institutions.* They do so under the circumstances detailed in the model. This finding closes the widening schism between microeconomic theories and empirical evidence started by Hardin (1968) and Olson (1965). To the extent that common goods are analytically similar to public goods (Bowles, 2004), it also makes some headway to reconcile other theories of collective action with the data.

Second, both CCA (10*) and NVD (11*) are essentially conditions on the within-period utility level at time of the expense in efforts relative to lifetime utility differences between 2 scenarios. In the current version of the model, this implies that a single-period subsidy to a carefully selected agent could be enough to ensure that she creates the self-governance mechanism and thus to shift the historical trajectory of the community towards sustainability. While this result must be qualified when we extend the model to a legitimacy-based budget for the Arbiter (where the Arbiter's power depends on popular support), this implies for example that a conditional lump-sum grant to a citizen/company/nation that is already motivated to solve the collective action problem can be enough to tilt the decision-maker towards action and establish a sustainable self-governance mechanism.

Another requirement for this second result to hold is that the Arbiter is reasonably designed and thus that the institution's targets are beneficial over the lifetime of the individual. If the targets are similar enough to what the agent would do if it was sole consumer of the common good, paying the transformation cost guarantees a solution to the collective action issue. Also, in reality, the transformation cost is likely to be spread over several periods and several agents (forming a coalition). Even if it cannot be demonstrated analytically in this first version of the model, we can expect that a strategic subsidy to carefully selected agents in specific periods are sufficient to ensure the creation of a reasonably designed Arbiter. The result mentioned above should therefore by qualified: *subsidies to ensure the Arbiter is created are small relative to the potential lifetime welfare gains from locking the community onto another more sustainable equilibrium path*, since the gains are multiplied by the number of agents and future periods while the cost is concentrated on a minority of agents for a few periods.

Third, from the argument of $u_i^{-1}(.)$ in (10*), I can deduce that the creation of the Arbiter is more and more likely as the good approaches exhaustion. Indeed, given the satiating consumption, the transformation cost and the discount factor, the period x at which the Arbiter is created determines how big the first summation term is: the later the x, the smaller the term. This term decreases faster with x than the negative summation term does, since $t^{**} > t^{*}$ when the Arbiter is designed to solve the tragedy of the commons. *The creation of the Arbiter is therefore delayed until the period where benefits from continued moderated consumption until* t^{**} *are greater than the opportunity cost of foregone consumption from x to* t^{*}.

4.2 Characteristics of successfully self-governed communities

While these general implications are already valuable, the analytical structure allows me to go even further and to observe the influence of each parameters on the outcome for a given community. Each community is characterized by the agents making it up. Assuming NVD$_i$ holds ex ante – i.e. the community is heterogeneous enough- the only condition for a community to experience the creation of an Arbiter is that CCA$_i$ holds for at least one agent. Whether a community solves the tragedy of the commons (or other collective action issues) is therefore dependent on how big the right-hand side (RHS) of (10*) is for a given agent i with exogenously given transformation cost e_i. The value of RHS depends on many factors that I explore here.

To reach a deeper level of analysis, I can re-express $p_{t,i}$ in (10*) as a function of exogenous parameters via (8*). I focus here on cases where the budget is binding ($\mu \sum_i^n (\hat{s}_{t,i} - s_{t,i}^*) > P$). Replacing $p_{t,i}$ in (10*), I obtain:

$$\begin{aligned} & e_i \leq \\ & \check{s}_i - u_i^{-1} \Bigg(\frac{1}{\beta_i^x} \Bigg[\sum_{t=x}^{t^*-1} \beta_i^t u_i(\check{s}_i) + \beta_i^{t^*} u_i\big(\hat{s}_{t^*,i}\big) + \sum_{t=t^*+1}^{t^{**}} \beta_i^t u_i(0) \\ & \quad - \sum_{t=x+1}^{t^{**}-1} \beta_i^t u_i \Bigg(\check{s}_i \\ & \quad - \frac{1}{n} \left[\frac{P}{\mu} + (n-1)\big(\check{s}_i - s_{t,i}^*\big) - \sum\nolimits_{j=1,j\neq i}^{n} \big(\check{s}_j - s_{t,j}^*\big) \right] \Bigg) \\ & \quad - \beta_i^{t^{**}} u_i\big(\hat{s}_{t^{**},i}\big) \Bigg] \Bigg) = RHS \end{aligned} \tag{28}$$

In discrete time, the fact that t^* and t^{**} are step functions (roundup functions) constrains the comparative statics. While I can use partial derivatives, the marginal change will always be required to happen holding t^* and t^{**} fixed i.e. "within the same step" of the roundup function. The extent to which this reduces the meaningfulness of the analysis depends on the mapping of t and T to the real world (the larger one period is relative to the aggregate T, the more meaningful). Note that this issue can be resolved by redeveloping the model in continuous time with a Hamiltonian dynamic optimization. t^* and t^{**} would then be derived as continuous functions of the same parameters as their discrete counterparts and a more elegant analysis of the CCA$_i$ could be performed via differentiation. I do not do this here.

Nevertheless, I can still make an instructive analysis. Let's consider the case where $r = 0$, since the growth rate does not qualitatively affect any decision in equilibrium[13]. Under these circumstances and assuming the target consumption levels are constant over time to have an explicit function for t^{**}, (4*) and (22) become:

$$t^* = roundup\left(\frac{S_0 - \sum_j^n \check{s}_j}{\sum_j^n \check{s}_j}\right) \tag{29}$$

[13] Indeed, since the strategic outcome is satiating consumption, the growth rate matters only when the Arbiter's targets are endogenous, which they are not in this analysis.

And

$$t^{**} = roundup\left(\frac{S_0 - \sum_i^n(\check{s}_i - \frac{1}{n}\left[\frac{P}{\mu} + (n-1)(\check{s}_i - s_i^*) - \sum_{j=1,j\neq i}^n(\check{s}_j - s_j^*)\right])}{\sum_i^n(\check{s}_i - \frac{1}{n}\left[\frac{P}{\mu} + (n-1)(\check{s}_i - s_i^*) - \sum_{j=1,j\neq i}^n(\check{s}_j - s_j^*)\right])}\right) \quad (30)$$

With these equations, I can evaluate the effect of a marginal change in different parameters on agent i's propensity to create the Arbiter and make preliminary predictions from these. Of course, not surprisingly, the smaller the transformation cost to agent i of creating the Arbiter, the more likely is CCA_i to hold everything else constant. Since e_i is purely exogenous, the model's usefulness comes from its predictions about other features of the agents and Arbiter, which I discuss next.

4.2.1 Agent's satiating consumption level $\check{s}_i$:

First, it is interesting to evaluate how the satiating consumption level of an agent provides information about its propensity to create the Arbiter. I show in Appendix 1.3 that evaluating whether big consumers of the common good are more or less likely to be the institutional entrepreneurs is a difficult task. The significant interactions between $\check{s}_i$ and the rest of the model makes the comparative statics highly sensitive to the initial value of $\check{s}_i$. Some examples under diverse reasonable specifications tend to indicate that the higher the level of satiating consumption, the more likely CCA is to hold conditional on satisfying $x < t^*$. This result is unfortunately not robust to more extreme specifications (cf. Appendix 1.3 for examples contradicting this result), even when the conditions for a determinate equilibrium are satisfied (i.e. $p_{t,i} \geq 0,\ 0 \leq x < t^*$, $0 \leq s^*_{t,i} \leq \check{s}_i$ and $0 < \beta_i < 1$).

4.2.2 Agent's discount factor β_i:

I can make more clear-cut predictions about who is going to create the Arbiter using the discount factor. Indeed, since neither t^* nor t^{**} depends on the discount factor, this prediction does not face the aforementioned caveat: it is globally valid. Taking the partial derivative of the RHS with respect to β_i I have:

$$\begin{aligned}\frac{\delta \text{RHS}}{\delta \beta_i} = -{u_i^{-1}}'(Z)\Bigg[&u_i(\check{s}_i)\sum_{t=1}^{t^*-x+1}(t^*-x-t){\beta_i}^{t^*-x-t-1} \\ &+ u_i\big(\hat{s}_{t^*,i}\big)(t^*-x)\beta_i^{t^*-x-1} + u_i(0)\sum_{t=t^*-x}^{t^{**}-x-1}(t+1){\beta_i}^{t} \\ &- u_i(\check{s}_i - p_i)\sum_{t=1}^{t^{**}-x-1}(t){\beta_i}^{t-1} \\ &- u_i\big(\hat{s}_{t^{**},i}\big)(t^{**}-x)\beta_i^{t^{**}-x-1}\Bigg] \end{aligned} \qquad (31)$$

With

$$
\begin{aligned}
Z = \sum_{t=x}^{t^{*}-1} \beta_i^{t-x} u_i(\check{s}_i) + \beta_i^{t^{*}-x} u_i\big(\hat{s}_{t^{*},i}\big) + \sum_{t=t^{*}+1}^{t^{**}} \beta_i^{t-x} u_i(0) \\
- \sum_{t=x+1}^{t^{**}-1} \beta_i^{t-x} u_i \Bigg(\check{s}_i \\
- \frac{1}{n}\left[\frac{P}{\mu} + (n-1)\big(\check{s}_i - s^{*}_{t,i}\big) - \sum\nolimits_{j=1, j\neq i}^{n} \big(\check{s}_j - s^{*}_{t,j}\big)\right]\Bigg) \\
- \beta_i^{t^{**}-x} u_i\big(\hat{s}_{t^{**},i}\big)
\end{aligned}
\tag{32}
$$

Since $0 \leq x < t^{*} \leq t^{**}$ in cases of interest and since $u_i^{-1'}(Z)$ is always positive (as the derivative of the inverse of a strictly increasing function), the effect of higher β_i will depend on the sign of the term in square brackets. Specifically, when this term is negative, the effect of a higher β_i on the propensity to create the Arbiter is positive. This term is negative when:

$$
\begin{aligned}
u_i(\check{s}_i - p_i) \sum_{t=1}^{t^{**}-x-1} (t){\beta_i}^{t-1} + u_i\big(\hat{s}_{t^{**},i}\big)(t^{**} - x)\beta_i^{t^{**}-x-1} \\
\geq u_i(\check{s}_i) \sum_{t=1}^{t^{*}-x+1} (t^{*} - x - t){\beta_i}^{t^{*}-x-t-1} \\
+ u_i\big(\hat{s}_{t^{*},i}\big)(t^{*} - x)\beta_i^{t^{*}-x-1} + u_i(0) \sum_{t=t^{*}-x}^{t^{**}-x-1} (t+1){\beta_i}^{t}
\end{aligned}
\tag{33}
$$

Therefore, CCA$_i$ is more likely to hold given a marginal increase in β_i when the marginal effect of β_i on lifetime utility after the creation of the Arbiter is bigger than its marginal effect on lifetime utility in the scenario where it is not created.

Given this conditionality on the effect, I use a reasonable specification to obtain a graph for $RHS(\beta_i)$ when $\mathrm{u_i(s_i) = ln\,(1 + s_i)}$, presented in figure 1. From figure 1, the greater the discount factor, the more likely CCA is to hold, conditional on $0 \leq x < t^* \leq t^{**}$, on having an Arbiter aiming to resolve the tragedy of the commons (i.e. $\check{s}_j \geq s_j^*$) and, crucially, on having reasonable starting value of β_i ($\beta_i \geq \sim 0.55$).

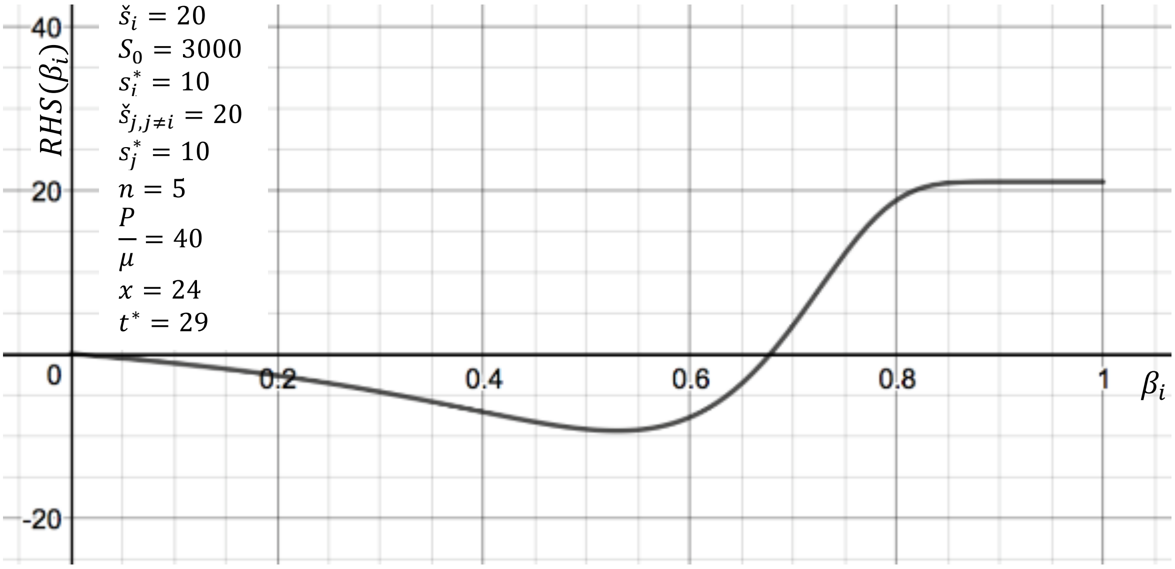

Figure 1: Effect of the discount factor on agent i's propensity to create the Arbiter, with $u_i(s_i) = ln\,(1 + s_i)$

When considering the entire distribution of discount factors in the population rather than a marginal change for a given agent, figure 1 also implies that agents with very high β_i are on average more likely than any other agent to create the Arbiter, assuming independence of the agent's β_i from other parameters. Notice the tendency towards a horizontal asymptote (at 21 with this specification), which implies that all else equal, *the maximum propensity to create the Arbiter is reached when β_i tends to 1.*

This result remains valid when varying the other parameters one at a time, so long as the parameters' values ensure a determinate equilibrium (i.e. subject to the conditions outlined above). This result is therefore robust to unidimensional variation in this baseline specification, at the very least, and for extreme variations. Therefore, *the model confirms the intuition that agents who are most future-oriented are most likely to act to solve the tragedy of the commons.*

4.2.3 Agent's target consumption s_i^*:

It is also interesting to see to what extent the constraint imposed by the Arbiter upon its creator affects the likelihood of creation. To do so, I study the effect of the target consumption set by the Arbiter for agent i on her propensity to create the Arbiter. The intuition is not clear as for how the target should affect the agent. If the target is too low (too frugal conditional on the enforcement budget), agent i's lifetime utility from the Arbiter starts to decrease because of lack of consumption relative to the level of the common good. If the target is too large, it is not worth paying the transformation cost since the Arbiter does not solve the issue and reduces the numbers of periods during which the good can be enjoyed. Fortunately, the condition $0 \leq s_i^* \leq \check{s}_i$ mutes this latter effect.

Figure 2 presents a graph of $RHS(s_i^*)$. Note that the function is not continuous even for $s_i^* < 40$, but the gaps are imperceptible. The discontinuity arises from the influence of s_i^* on t^{**}. Since each "piece" of the function is strictly increasing and concave, we see that the agent prefers higher targets given a marginal increase in the target does not reduce the number of periods during which she can consume. This explains the upward slope within each step.

On the domain $0 \leq s_i^* \leq \check{s}_i = 20$, the function is a succession of increasing segments, with maximum on this domain at $s_i^* = 20$. This implies that the agent's maximum propensity to create the Arbiter is constrained to be at $s_i^* = \check{s}_i$. *Agent i is more likely to create the Arbiter if agent i's allocated target is equal to her satiating consumption.*

This result is supported even when varying the other parameters one at a time, so long as the parameters' values ensure a determinate equilibrium (i.e. subject to the conditions outlined above). This result is therefore robust to unidimensional change to this specification, at the very least, and for extreme unidimensional changes as well.

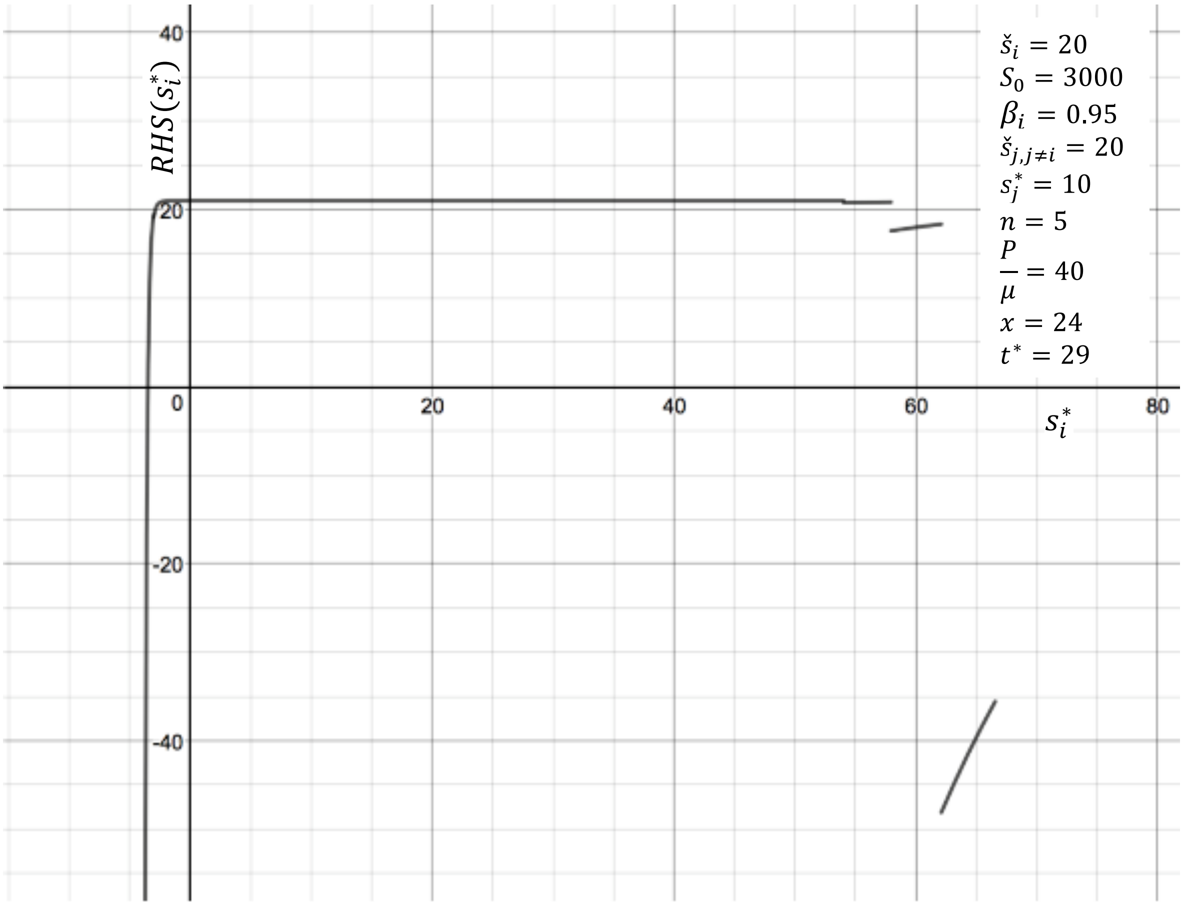

Figure 2: Effect of agent i's Arbiter-set consumption target on agent i's propensity to create the Arbiter, with $u_i(s_i) = ln\,(1 + s_i)$

4.2.4 Other agents' target consumption s_j^*:

After looking at the agent's own target, the natural question is to evaluate how the targets set for other agents affect the willingness to create the Arbiter of a given agent. Because of how the others' targets affect the aggregate consumption under the Arbiter's regime, it affects the number of periods during which the agent creating the Arbiter enjoys the good. Furthermore, given a certain budget, the smaller the targets for the other agents, the less intense will the punishment be on the agent creating the Arbiter. Therefore, the RHS should decrease as the targets set for the other agents increase.

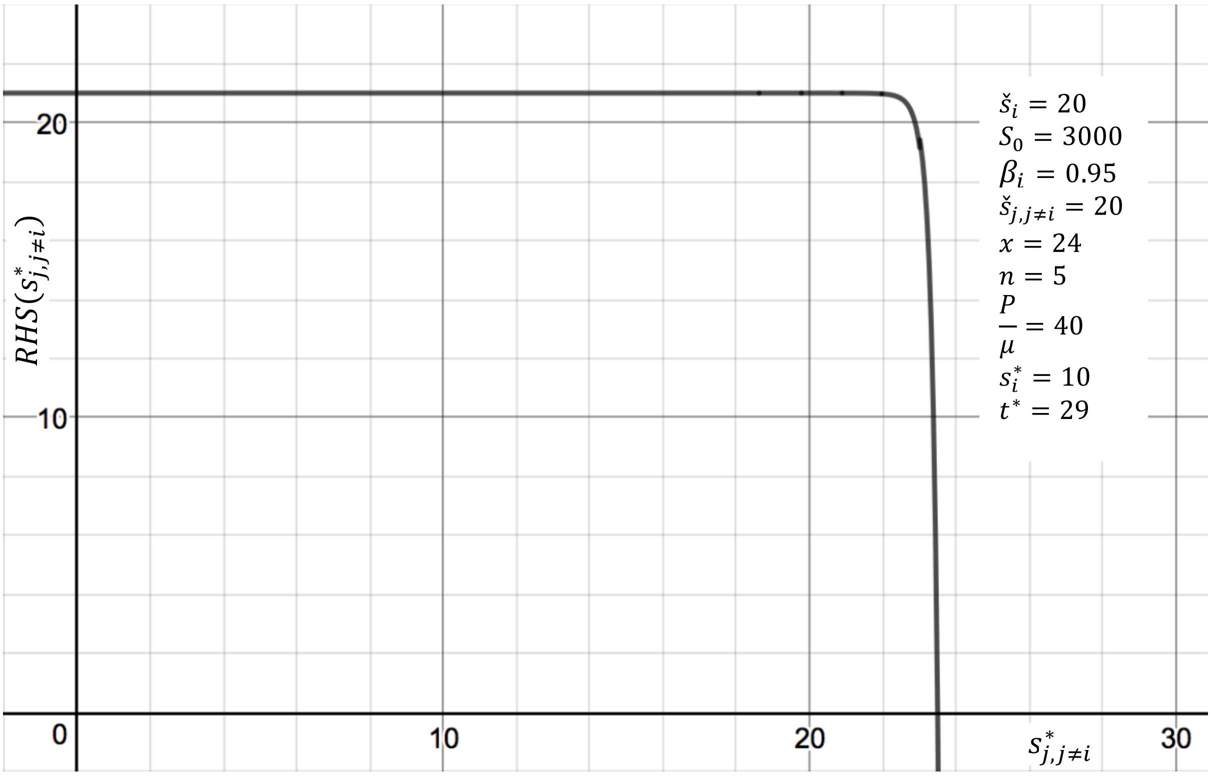

Figure 3: Effect of other agents' targets on agent i's propensity to create the Arbiter, with $u_i(s_i) = ln\,(1 + s_i)$

Figure 3 confirms this reasoning. Note again that the function is discontinuous. For simplicity, I have also assumed that all the agents besides agent i are given the same targets. As the targets are more lenient, it is less useful for agent i to create the Arbiter, since t^{**} occurs earlier. Since the model is designed for cases where $0 \leq s_i^* \leq \check{s}_i$, I provide figure 4 to zoom in on targets that induce greater sustainability. Although it is not perceptible on figure 3, figure 4 shows there is a clear preference for more constraining targets for others. *An agent's propensity to create the Arbiter decreases when the Arbiter is too lenient for others.*

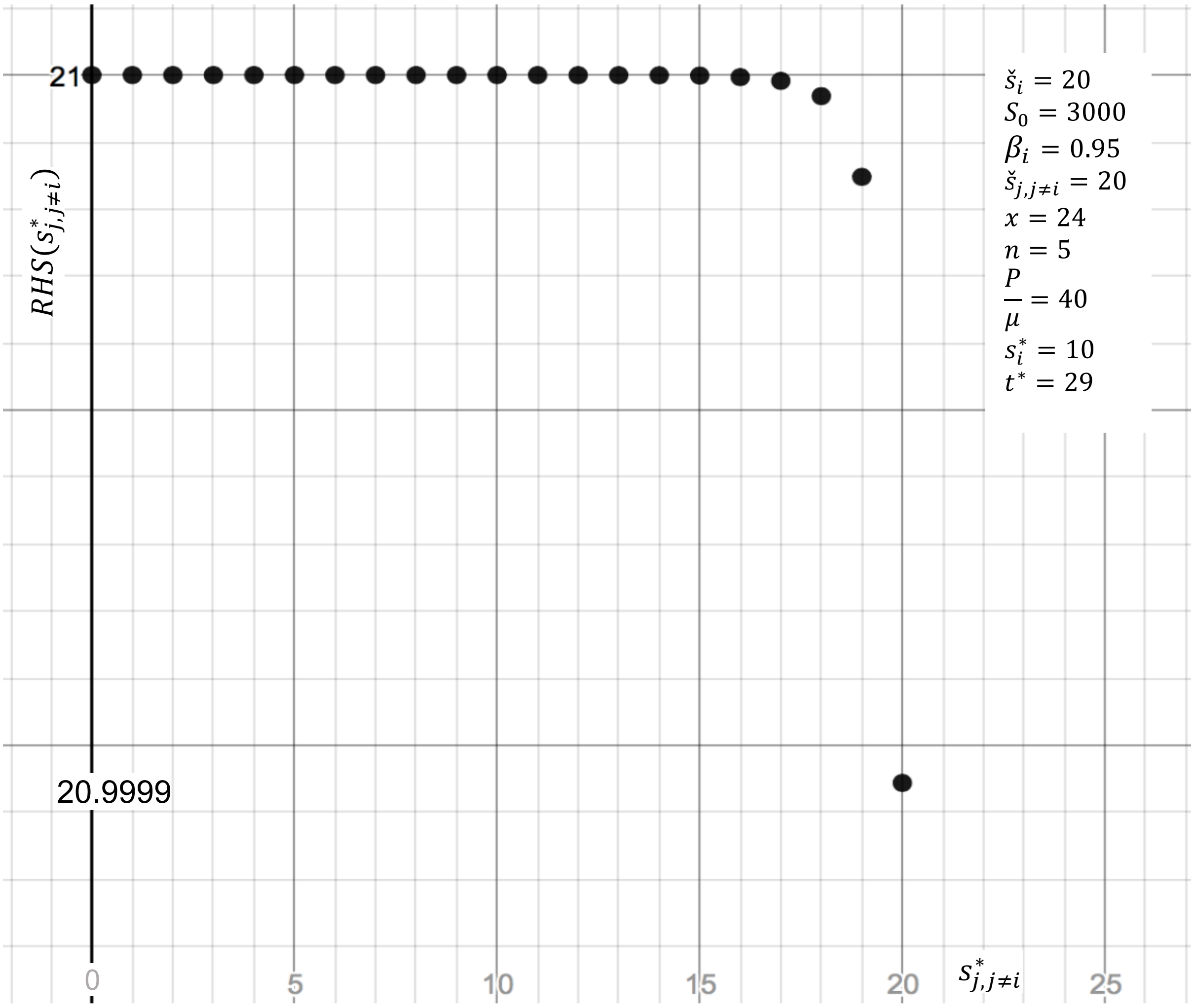

Figure 4: Effect of other agents' targets on agent i's propensity to create the Arbiter (graph truncated at $s^*_{j,j\neq i} > 20$*), with* $u_i(s_i) = ln\ (1 + s_i)$

This result is supported even when varying the other parameters one at a time, so long as the parameters' values ensure a determinate equilibrium (i.e. subject to the conditions outlined above). This result is therefore robust to unidimensional change to this specification, at the very least, and for extreme unidimensional changes too. Of course, *I expect this result to disappear when introducing a legitimacy-based Arbiter, where popular support matters for power.*

4.2.5 Timing of the creation of the Arbiter x:

A crucial insight from the model is a prediction about what is the most propitious time for creating the Arbiter. Effectively, by consider x as a parameter for each agent, I can determine when the agent is most inclined to create the Arbiter and therefore solve the collective action problem given his preferences. I therefore plot the value of the RHS as a function of time at which the decision is made. A single maximum can be identified (here, at $x = 25$ though it is difficult to see). Figure 5 shows *the propensity to create the Arbiter peaks at a specific period close enough to but before the last opportunity to do so.* Given the Arbiter's targets are exogenous, acting too early would indeed be suboptimal since the stock is not close enough to exhaustion to justify instituting an exogenous constraint. However, the agent prefers not to wait until the very last period to act (i.e. $x = 28$ is *not* the maximum) so that there is enough of the stock left to smooth consumption over a longer time window $t^{**} - t^{*}$ under the Arbiter's regime. This formalizes the third general implication of section 4.1.

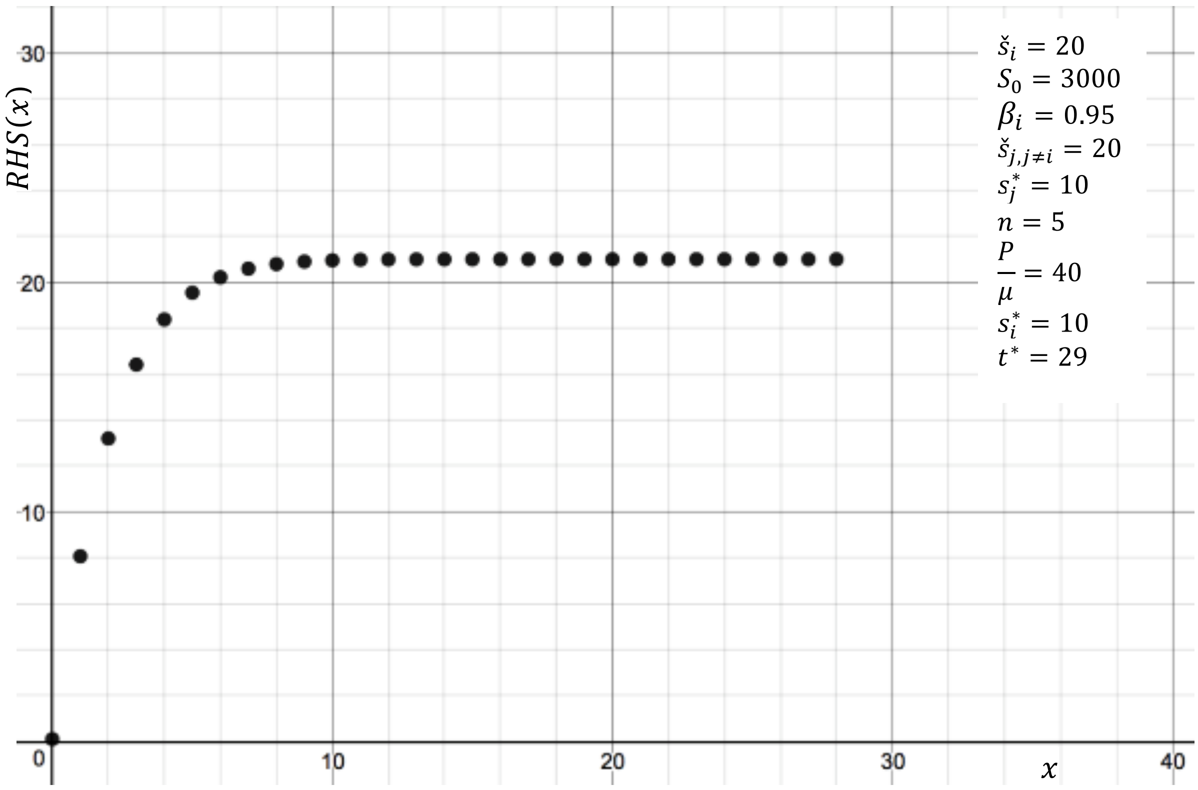

Figure 5: Effect of the timing of the creation on agent i's propensity to create the Arbiter, with $u_i(s_i) = ln\ (1 + s_i)$

This result remains valid when varying the other parameters one at a time, so long as the parameters' values ensure a determinate equilibrium (i.e. subject to the conditions outlined above). This result is therefore robust to unidimensional changes in the specification, at the very least, and to extreme unidimensional changes as well. Determining this maximum analytically is an issue left for further research.

4.2.6 Number of agents in the community n*:*

To the extent that the number of agents drives up aggregate consumption, it makes t^* and t^{**} happen sooner. For any given x, the agent has therefore stronger incentives to act as n grows, so that it can force everybody to restrain before the good is exhausted. However, beyond a certain n, a greater number of agents "dilutes" the given power of the Arbiter ever more, so that it becomes less and less useful in altering the transition path, relative to the transformation cost. There is therefore a trade-off between these two effects, which would imply that there exists an optimal n for agent i's propensity to create.

Even assuming NVD holds ex ante, given that n affects both t^* and t^{**} directly, as well as indirectly via the Arbiter's punishment decision, there are interaction effects between n and the other parameters. This notably leads to thresholds and oscillations in the propensity to create the Arbiter. I have confirmed this by trying different specifications. While the reasoning in the previous paragraph remains valid across specifications, I cannot characterize the single optimal n further than by inferring it exists. Here again, it would be valuable to redevelop the model in continuous time to make the comparative statics of n more tractable.

4.2.7 Effective budget of the Arbiter $^{P}/_{\mu}$*:*

Is it worthwhile to create a weak Arbiter, with little effective power to sanction? This an important question in Politics and International Law. Figure 6 displays the evolution of the propensity to create the Arbiter as a function of the effective budget. Not surprisingly, it mirrors the effect of the target consumption set by the Arbiter for other agents (figure 3). Since it enters the punishment function with a sign opposite to the target consumption, it is to be expected. The function is not continuous. It also displays a single maximum so long as the parameters satisfy the conditions described above such that a determinate equilibrium exists. Note that although I have included values of the budget beyond 50 to illustrate the global (dis)similarity with the target consumptions, the analysis here is appropriate only for values of the budget up to 50, because the budget must be binding in this specification. Therefore, conditional on having a binding budget, *an agent is more likely to create an Arbiter with a bigger budget.*

This result is supported even when varying the other parameters one at a time, so long as the parameters' values ensure a determinate equilibrium (i.e. subject to the conditions outlined above). This result is therefore robust to unidimensional change to this specification, at the very least, and for extreme unidimensional changes too.

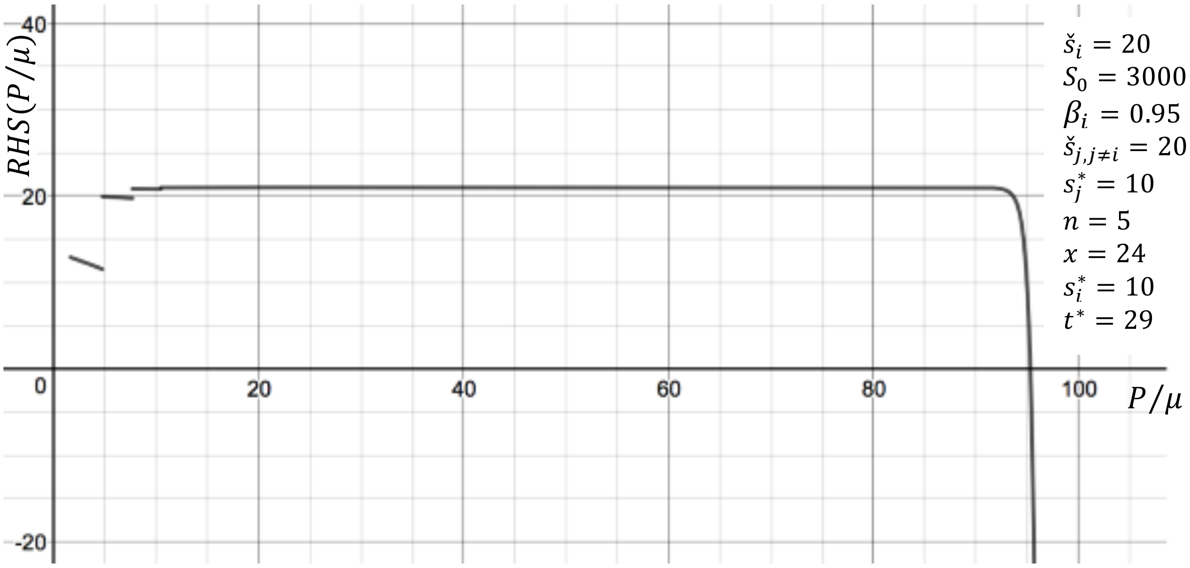

Figure 6: Effect of the Arbiter's effective budget on agent i's propensity to create the Arbiter, with $u_i(s_i) = ln\,(1 + s_i)$

4.2.8 Starting stock of the common good S_0:

Finally, determining whether there is an optimal starting stock for a given community could be interesting to explain variance across communities. Is a community with a greater natural endowment more likely to create an Arbiter to protect it, everything else held constant? The answer from this model seems to be "not necessarily", as figure 7 illustrates. For any given period where the agent decides whether to create the Arbiter, there is a (broad) global middle-ground of optimal values of S_0. Indeed, it would be meaningless to pay now for restraining the community from consuming a stock that is currently so large that exhaustion occurs only far in the future. Likewise, it would be inefficient for an agent to pay the transformation cost if the stock is already so low that even the Arbiter cannot sustain the community's consumption much longer.

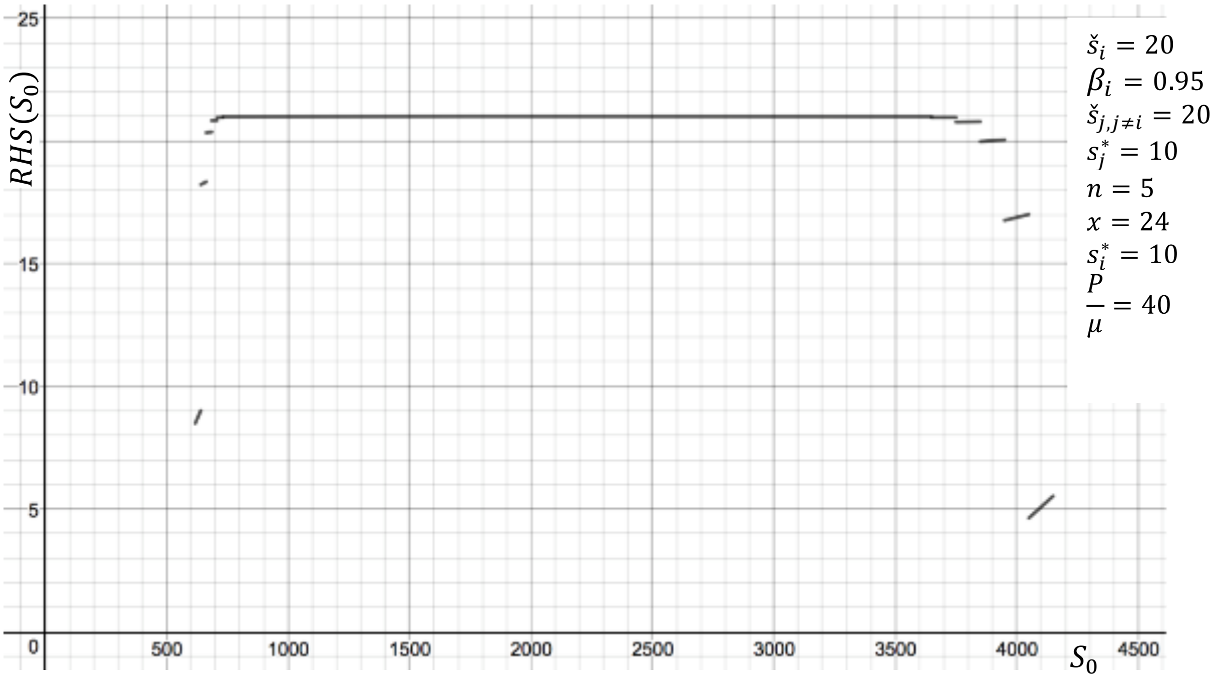

Figure 7: Effect of a greater starting stock on a community's propensity to create the Arbiter, with $u_i(s_i) = ln\,(1 + s_i)$

This confirms the result obtained for the propitious timing of the creation of the Arbiter: not too early before exhaustion (S_0 cannot be too large), but not too late neither so that there is enough common good left to spread over several periods in the future, ensuring it is worth it to create the Arbiter. This is also robust to unidimensional change to the specification, including extreme changes.

4.3 Aggregate welfare impact of an Arbiter

I now turn to the effect of a given Arbiter on the aggregate welfare of the community. Naturally, since the creation of the Arbiter only depends on having one individual willing to create it and since individuals are heterogeneous, the Arbiter does not systematically improve aggregate welfare. Although it does offset the aggregate welfare loss from reducing the impact of strategic behaviours, the new consumption targets might be so misaligned with the other agents' preferences that there is a net decrease in lifetime welfare.

Evaluating aggregate welfare requires making a value choice on the social welfare function. Here, I assume a plain utilitarian social welfare function. The sum of utility levels generated in the community and perceived by agents is what matters, implicitly assuming inequality across individuals and/or periods has no effect on welfare. Since the creation of the Arbiter depends on CCA_i that itself depends on the other parameters, I effectively compute the aggregate welfare change from having a marginal change in e_i such that a first CCA_i holds in the community (that is, a marginal change in the model that results in a change of equilibrium paths).

Denoting aggregate welfare in the community by W, and recognizing that the welfare difference occurs only from x to t^{**}, I have:

$$
\begin{aligned}
\Delta W =\ & W_{Arbiter} - W_{Baseline} \\
=\ & \left\{ \sum_{j \neq i}^{n} \beta_j^x u_j(\check{s}_j) + \beta_i^x u_i(\check{s}_i - e_i) \right. \\
& \qquad \left. + \sum_{j}^{n} \left[\sum_{t=x+1}^{t^{**}-1} \beta_j^t u_j\left(\check{s}_j - p_{t,j}\right) + \beta_j^{t^{**}} u_j\left(\hat{s}_{t^{**},j} - p_{t^{**},j}\right) \right] \right\} \\
& \qquad - \left\{ \sum_{j}^{n} \left[\sum_{t=x}^{t^{*}-1} \beta_j^t u_j\left(\check{s}_j\right) + \beta_j^{t^{*}} u_j\left(\hat{s}_{t^{*},j}\right) \right. \right. \\
& \qquad \left. \left. + \sum_{t=t^{*}+1}^{t^{**}} \beta_j^t u_j(0) \right] \right\}
\end{aligned}
\tag{34}
$$

This expression boils down to:

$$
\begin{aligned}
\Delta W =\ & \beta_i^x [u_i(\check{s}_i - e_i) - u_i(\check{s}_i)] \\
& + \sum_{j}^{n} \left[\sum_{t=x+1}^{t^{*}-1} \beta_j^t \left[u_j\left(\check{s}_j - p_{t,j}\right) - u_j\left(\check{s}_j\right)\right] \right. \\
& + \beta_j^{t^{*}} \left[u_j\left(\check{s}_j - p_{t^{*},j}\right) - u_j\left(\hat{s}_{t^{*},j}\right)\right] \\
& + \sum_{t=t^{*}+1}^{t^{**}-1} \beta_j^t \left[u_j\left(\check{s}_j - p_{t,j}\right) - u_j(0)\right] \\
& \left. + \beta_j^{t^{**}} \left[u_j\left(\hat{s}_{t^{**},j} - p_{t^{**},j}\right) - u_j(0)\right] \right]
\end{aligned}
\tag{35}
$$

Given the assumptions about $u_i(.)$, the sign of this expression depends on the discount factors in the population and the targets set by the Arbiter (since they determine the difference between t^{**} and t^{*} as well as the within-period utility levels when the Arbiter is created). The within-period utility differences associated with periods x to $t^{*} - 1$ (and potentially to t^{*}, depending on the sign of $\check{s}_j - p_{t^*,j} - \hat{s}_{t^*,j}$) are negative while the ones associated with periods t^* (potentially $t^* + 1$) to t^{**} are positive. This illustrates the short-term sacrifice required to enjoy the long-term benefits from the good.

(35) makes clear that the Arbiter's sanctions and, implicitly, the targets are crucial to determine the welfare impact. Indeed, the welfare differential roughly reduces to a discounted series of aggregate comparisons between the utility from $\check{s}_j - p_{t,j}$ with $\check{s}_j$ and 0. The closest the punishment pushes an agent towards its naïve consumption path, the closest to optimal that agent's utility for that period will be, considering the dynamic trade-off. If the punishments are arbitrary, however, the agents are forced onto an arbitrary consumption path, which in general cannot achieve results close to optimality. If the punishment is arbitrarily lenient for one agent but not for the other, that agent might achieve higher utility than in her naïve scenario, but at the cost of having another agent receiving sanctions so strict that she has lower lifetime utility than in the baseline scenario. This highlights some distributional implications of the Arbiter that I do not have the space to explore in this thesis.

Nevertheless, there is one analytical result about welfare that is interesting to explore here. Empirically, we are unlikely to observe both the situation where the Arbiter is not created and its counterfactual for a given community. It is therefore interesting to extract as much information as possible from (35) without the need for data on both situations. Given the terms of this expression are all differences between two different levels of the same function, I can use the mean value theorem to re-express (35), so long as these differences are different from 0. The resulting expression is:

$$\begin{aligned}\Delta W = {} & \beta_i^x[-e_i u_i'(A_i)] \\ & + \sum_j^n \left[\sum_{t=x+1}^{t^*-1} \beta_j^t[-p_{t,j} u_j'(B_j)] \right. \\ & + \beta_j^{t^*}\left[\left(\check{s}_j - p_{t^*,j} - \hat{s}_{t^*,j}\right) u_j'(C_j)\right] \\ & + \sum_{t=t^*+1}^{t^{**}-1} \beta_j^t\left[\left(\check{s}_j - p_{t,j}\right) u_j'(D_j)\right] \\ & \left. + \beta_j^{t^{**}}\left[\left(\hat{s}_{t^{**},j} - p_{t^{**},j}\right) u_j'(E_j)\right] \right] \end{aligned} \tag{36}$$

$$with$$

$$\check{s}_i - e_i < A_i < \check{s}_i; \quad \check{s}_j - p_{t,j} < B_j < \check{s}_j; \quad 0 < D_j < \check{s}_j - p_{t,j}; \\ 0 < E_j < \hat{s}_{t^{**},j} - p_{t^{**},j}$$

$$and$$

$$\begin{aligned} & \check{s}_j - p_{t^*,j} < C_j < \hat{s}_{t^*,j} \quad if\ \check{s}_j - p_{t^*,j} < \hat{s}_{t^*,j} \\ or \quad & \hat{s}_{t^*,j} < C_j < \check{s}_j - p_{t^*,j} \quad if\ \hat{s}_{t^*,j} < \check{s}_j - p_{t^*,j} \end{aligned}$$

Owing to the strict concavity and continuity of $u_j'(.)$, the mean value theorem enables me to derive boundaries on $u_j'(.)$. From these, I can obtain the following boundaries on ΔW:

$$\begin{aligned}\Delta W > \beta_i^x[-e_i u_i'(\check{s}_i - e_i)] \\ + \sum_j^n \Bigg[\sum_{t=x+1}^{t^*-1} \beta_j^t[-p_{t,j} u_j'(\check{s}_j - p_{t,j})] \\ + \beta_j^{t^*}[(\check{s}_j - p_{t^*,j} - \hat{s}_{t^*,j}) u_j'(\check{s}_j - p_{t^*,j})] \\ + \sum_{t=t^*+1}^{t^{**}-1} \beta_j^t[(\check{s}_j - p_{t,j}) u_j'(\check{s}_j - p_{t,j})] \\ + \beta_j^{t^{**}}[(\hat{s}_{t^{**},j} - p_{t^{**},j}) u_j'(\hat{s}_{t^{**},j} - p_{t^{**},j})] \Bigg]\end{aligned} \tag{37}$$

$$\begin{aligned}\Delta W < \beta_i^x[-e_i u_i'(\check{s}_i)] \\ + \sum_j^n \Bigg[\sum_{t=x+1}^{t^*-1} \beta_j^t[-p_{t,j} u_j'(\check{s}_j)] \\ + \beta_j^{t^*}[(\check{s}_j - p_{t^*,j} - \hat{s}_{t^*,j}) u_j'(\hat{s}_{t^*,j})] \\ + \sum_{t=t^*+1}^{t^{**}-1} \beta_j^t[(\check{s}_j - p_{t,j}) u_j'(0)] \\ + \beta_j^{t^{**}}[(\hat{s}_{t^{**},j} - p_{t^{**},j}) u_j'(0)] \Bigg]\end{aligned} \tag{38}$$

The added value of (37) is significant. Indeed, this lower bound on ΔW can be computed with much more accessible information. By converting an expression with utility levels into an expression with marginal utility at these levels, it simplifies a potential empirical estimation. Indeed, marginal utility is more easily observed (via relative prices or willingness to pay) than cardinal utility levels. More importantly, the lower boundary *does not require an estimation of the disutility to agents of not having the common good anymore.* This saves the researcher from the tedious task of e.g. estimating the marginal disutility from not having fossil fuels or international fisheries (the markets for which would no longer exist) or guessing nations' willingness to pay for no climate change in a world tormented by a runaway climate change scenario. I have included (38) for the sake of completeness.

4.4 Testing the model's predictions

As mentioned in section 2, I evaluate my model' validity against Ostrom's facts. These facts are the outcome of the research agenda she steered. They concisely summarize the results of dozens of case studies on self-governed CPR communities across the world. In her empirical research, Ostrom (2000) has highlighted 8 "design principles" for the long-term success of self-organized institutions for collective action. With the terminology used in the current model, these principles are (Ostrom, 2000):

1. *Clear boundary rules*: all agents and the Arbiter know who has access to the common resource and who is under the Arbiter's authority

2. *The Arbiter's targets are*:
    a. *Constraining*: the rules restrict consumption
    b. *Proportional*: the Arbiter sets targets by taking the agent-specific preferences into account

    c. *Tailored*: the rules take the stock level into account.

3. *Democratic Arbiter*: all or at least most agents affected by the Arbiter's regime can influence the rules set by the Arbiter.

4. *Arbiter's accountability*: the agents select their own Arbiter, accountable to the community, monitoring agents' behaviour and the common good's level.

5. *Graduated sanctions*: the punishment of an agent depends on the extent to which it deviates from the rules

6. *Conflict resolution*: there is a low-cost mechanism to resolve conflicts in the community

7. *Recognition by higher authority*: the community's right to create an Arbiter is recognized by a central authority

8. *Multiple layers of self-organization*: in larger communities, the structure Arbiter-Agents is extended, with Arbiters at the lowest level of hierarchy being monitored by Arbiters at a higher level.

Additionally, Ostrom (2000) observed 2 conditions that enable the emergence of a self-governance mechanism:

9. *Incentives to cooperate*: there exist some agents in the community willing to restrict their consumption if almost everybody else reciprocate.

10. *Institutional entrepreneur*: there is an institutional entrepreneur ready to initiate the creation of the Arbiter.

The goal of a theory of CPR management is to be consistent with these 10 facts. If the model and its equilibrium are consistent with these empirical facts, it is a valid theory of how anarchic communities self-organize to solve common goods issues. If, besides, its assumptions and microeconomic foundations are consistent with observed behaviours, then the theory gains plausibility among the set of theories that could generate these facts. Of course, the Ostrom's facts are rich and nuanced. Contrary to Kaldor's facts (Kaldor, 1957), Ostrom's facts were certainly not developed with the intention of eventually testing them in structural models. I therefore did not attempt to satisfy all 10 facts, since I need analytical tractability for the model to be useful in Economics.

The power of these facts over, say, my own case studies for international cooperation (available upon request), is that they are the conclusion of dozens of careful studies of self-governed CPR communities by scholars worldwide over the past 50 years. While the model remains agnostic about some of these facts, I prefer to list them anyway to highlight how much progress there remains to be made to develop a general theory of self-organization for collective action problems. I am nevertheless happy to list a few successes. I discuss below how the model compares to these facts and how it could be improved:

1. Clear boundary rules: I have specified the set of agents in the community as well as the set of agents under the Arbiter's authority. This means the model is consistent by assumption, so that the model cannot be applied to situations where clear boundary rules do not exist. This is therefore neither a success nor a failure of the model.

2. The Arbiter's targets are:
    a. Constraining: the rules are indeed constraining to the extent that the Arbiter has a positive budget. The model allows to evaluate the statement and its counterfactual (i.e. setting targets that are not constraining or that cannot be enforced). I have shown that the Arbiter is more likely to be created if its targets constrain the community (cf. figure 4), because only then is it worth paying the transformation cost. This is the first success of this model.

b. Proportional: the model can accommodate both proportional and non-proportional targets. I however did not cover whether proportionality was making it more likely for the Arbiter to exist. Intuitively, the current version of the model is agnostic about proportionality: since the agent creating the Arbiter does not care about others', she is not more likely to establish the Arbiter when targets are proportional. This is a first failure of this version of the model, though I expect a legitimacy-based Arbiter (discussed below) to resolve this inconsistency with the evidence.

c. Tailored: the model can accommodate both tailored and arbitrary targets, depending on the Arbiter's agenda-setting mechanism. An earlier version of this thesis expanded on endogenous targets and highlighted how tailoring to local stock was necessary for the Arbiter's targets to be more in-line with the naïve agent's ideal consumption (the socially optimal one). Naturally, an agent will be more willing to pay for the Arbiter only if the targets are indeed bringing her closer to her naïve consumption path, which itself depends on the level of the common good. This is therefore a second success of the model, that I unfortunately did not have space to discuss extensively.

3. Democratic Arbiter: the current version of the model is far from accommodating the level of democracy of the Arbiter. I assume a given design available to the community. An extension providing a form of bargaining over agenda-setting could be helpful. A legitimacy-based budget (discussed below) that would effectively constrain what agenda of the Arbiter can ever be agreed upon could also work. This is a second failure of the model.

4. Arbiter's accountability: to a certain extent, with the full-information world assumed here, accountability to the community is synonymous of democracy. Indeed, by assumption the Arbiter monitors both the agents' individual consumption and the evolution of the good (if targets are endogenous, as explained above). Since the counterfactual cannot be tested, this is neither a failure nor a success of the model.

5. Graduated sanctions: a key contribution of the model is to embed an Arbiter as a separate agent with its own objective and decision-making given its environment. This allows the Arbiter to tailor its punishments to sanction proportionally to the deviation from the targets. Of course, punishments could be made exogenous and would allow me to compare whether a given Arbiter is more likely to be established with fixed sanctions. Fixed punishments drive the self-governed community's path

away from each agent's ideal naïve path, so that it is less likely to be created. This is therefore a third success of the model.

6. Conflict resolution: the problem of conflict is assumed away from the model because of rationality and full information. Modelling such complex behaviour as bilateral or multilateral conflicts among agents, and upgrading the Arbiter with a court system is beyond the scope of this paper. This is therefore neither a failure nor a success of the model.

7. Recognition by higher authority: the presence of a central authority is also assumed away from the model by anarchy. Again, developing a model of central authority-community interactions is beyond the scope of this paper. This is therefore neither a failure nor a success of the model.

8. Multiple layers of self-organization: the concept of "larger" communities implies a geographic coverage or at least some form of scale through which it could be more efficient to set up Arbiters with authority over a subset of the community, themselves monitored by an Arbiter. This is also beyond the scope of the current version of the model, though it seems that the same structure could be re-iterated at the Arbiter level.

The model remains agnostic regarding this prediction, so this is neither a failure nor a success.

9. Incentives to cooperate: my analysis of the naïve agent's behaviour shows there clearly are first-order incentives to cooperate. However, the perfect rationality assumption imposes a collapse into the tragedy of the commons that cannot be reconciled with conditional reciprocation that Elinor Ostrom probably means by this fact. Indeed, throughout the paper, the harsh truth is that the agents cannot cooperate, but can just pay for an Arbiter to force them onto a path similar to successful cooperation. To a certain extent, this is therefore a third failure of the model.

10. Institutional entrepreneur: the model accommodates this prediction. If there is no institutional entrepreneur willing to pay the cost, the Arbiter cannot be created. The model goes beyond this fact and even offers some preliminary insights about who is likely to become institutional entrepreneur: somebody future-oriented, who would not suffer from the Arbiter's targets. This is therefore the fourth and most resounding success of the model.

Overall, the model provides a good starting point for a theory of self-organization, with 4 out of 7 evaluated facts successfully integrated. I expect that extending the model with a feedback mechanism from the community to the Arbiter will help successfully integrate 2 additional facts. In the next section, I explain several difficulties I have faced when attempting to model such a feedback mechanism via a legitimacy-based Arbiter.

5. A difficult extension: Legitimacy

We see from the reality check that the model's greatest structural weakness is the absence of feedback mechanisms from other agents in the community to the Arbiter. Effectively, an outlying agent could create a corrupt Arbiter even if all the other agents were suffering from this decision. This is unrealistic. Even the harshest dictatorial institution still relies on some support base or at least has some means to extract support from the community. Indeed, the population's approval is the reason why certain local courts are deemed as legitimate and others are not. The idea that formal institutions require "legitimacy" to be effective has been extensively studied in the International Relations and Politics literature and remains controversial (Dogan, 2002 for a review of the concept).

There are several artificial ways to introduce legitimacy[14]. A simple feedback loop could be a "vote of confidence" on the Arbiter, whereby if less than a certain proportion of the community favour the creation of the suggested Arbiter at a certain period, the Arbiter cannot be created that period. Similarly, I could inspire myself from Bueno de Mesquita et al.'s selectorate model and assume that if the coalition of supporters falls below a certain threshold, the Arbiter is temporarily deposed (Bueno de Mesquita et al., 2005, Chap. 3; Morrow et al., 2008). However, both approaches require assuming an arbitrary threshold. I have already had to assume an exogenous e_i to close the model and it would not be in good taste to add a second empirically obscure parameter to a model at its first iteration[15].

To develop an endogenous legitimacy-based Arbiter, I would ideally define the power of the Arbiter as an increasing function of the ratio of agents who prefer the Arbiter to exist. An agent's preference over whether to have an Arbiter itself depends on the consumption agenda that the Arbiter sets and its enforcement. I would therefore expect a trade-off between wanting the Arbiter to solve the issue, while being individually allowed to consume as much as possible. The set of potential consumption agendas for an effective Arbiter would therefore be constrained by whether a given agenda has enough support in the community

[14] Note that in this full information context with utility functions as defined above, Scharpf (1997)'s traditional distinction between input and output legitimacy is irrelevant.

[15] Also, too keen to develop the ideal budget function, it occurred to me too late that a simple vote of confidence could be operationalized. I will therefore work on this aspect for the final version of the paper.

to be enforced. If the Arbiter is welfare-maximizing and fair, it is more likely to obtain the support of a broader base and thus to be more effective. An institutional entrepreneur would therefore face the trade-off between picking an Arbiter design that is popular in the community (and thus effective in enforcing its agenda) and a design that particularly advantages her.

There are however a couple of analytical issues with integrating a legitimacy-based budget in the model. First, if the support base varies over time, t^{**} cannot be expressed as a simple function of the starting stock and aggregate consumption. It must be derived from all the other parameters. However, whether to support the Arbiter also depends on t^{**}. So, there seems to be an indeterminacy lurking on the horizon, which can only be resolved by developing a dynamic model for t^{**} and finding its equilibrium. Second, there is now a feedback loop: effective consumption affects the support base, which affects power, which affects sanctions which in turn affect effective consumption. This adds an additional layer of complexity to the model, that can only be circumvented by assuming the Arbiter sets an agenda without considering the effect it has on its support base.

More problematic is the fact that a support base comes with its own lot of strategic interactions. Whether to support the Arbiter depends on the support it already has. Even though the first order incentives of an agent might be to support the presence of an Arbiter, this agent might have second order incentives to withhold her support given the existing support base, so that the Arbiter cannot punish her as effectively. Considering the presence of an Arbiter as a form of public good, we could use game theoretical insights about coalition-building for public good provision (cf. Ray & Vohra (2015) for an introduction).

By contrast with the volunteer's dilemma for creating the Arbiter, I cannot derive a simple condition to isolate cases where this coalition-building issue does not occur. Using the assumption that agents do not behave strategically for coalition-building would break the rationality assumption that I have managed to maintain throughout. As with the volunteer's dilemma, I did not manage to integrate the findings from games of coalition formation for public goods. Further attempts using the open membership models by Yi and Shin (1995) are necessary, relying on Yi (1997) for stability properties and Belleflamme (2000) for the impact of heterogeneity.

This lack of legitimacy of the Arbiter is problematic, as it would improve the results' consistency with reality, especially in the international community. I indeed expect qualitative changes to the model's equilibrium from having the Arbiter's power affected by whether it is legitimate in the community: whether the Arbiter's targets are popular matters. It would also allow to extend the applicability of the model to the growth literature, where formal institutions' inclusiveness has recently been put forward as key precondition for growth (Acemoglu, Johnson & Robinson, 2005). A structural approach to inclusiveness could improve compatibility with existing growth models. A successful integration of legitimacy in the present model would therefore help to answer such fundamental questions as why of all the institutional designs possible some communities select a corrupt Arbiter like a totalitarian local government or a welfare-maximizing one like a council respecting citizens' preferences?

This failure of the current model is why I believe there are productive gains from merging this collective action model with models of social innovation based on evolutionary game theory and models of network. Indeed, contrary to the model presented here, these network-based models describe social feedback on new rules (Young, 2015).

6. Conclusion

I have developed a new model of the tragedy of the commons and examined an endogenous solution to the issue. I have shown that even in anarchic communities of rational agents with purely egoistic preferences, some agents have incentives to initiate the costly creation of a self-governance mechanism. The result relies on having agents with satiation points, so that the tragedy of the commons occurs over several periods; heterogeneous preferences so that the agents overcome the volunteer's dilemma and, crucially, on giving agents the ability to establish the Arbiter in exchange of a transformation cost.

The thesis provides a new explanation to the self-governance of common-pool resources observed worldwide throughout history and, more fundamentally, a new approach to study collective action problems and the evolution of institutions. Crucially, this does not rely on social preferences or irrationality. The theory also makes empirical predictions consistent with some of the stylized facts summarized by Ostrom (2000). It has made a concrete contribution to this field by providing a tractable mechanism for creating formal institutions. In the broader context of this line of research, it has effectively made the "experimentation" with rules in Young (2011) endogenous.

This was possible only by taking a radically different approach to collective action problems. Indeed, already in the introduction to this piece I made my point of view clear by claiming that global collective action problems are solvable. In section 2, I then elaborated on the evolution of the theoretical and empirical literature about collective action problems. In particular, I have highlighted the traditional theoretical result of a failure to cooperate while empiricists have repeatedly found evidence of the contrary. The core contribution of this paper is section 3, where I back up my claims with a model that describes the conditions under which a cooperation mechanism is created among rational, asocial and memoryless agents.

The thesis also evaluates this model and discusses its implications in section 4. Although the model is exploratory, it extracts some insights: broad implications for self-organization under anarchy, specific predictions about the propensity of a community to establish a self-governance mechanism and basic welfare implications. Additionally, section 4.4 shows how the model compares to the existing evidence. This empirical evaluation of the model highlights the need for integrating legitimacy in the Arbiter. Section 5 therefore outlines a strategy to do so, but explains the analytical pitfalls that prevent me from implementing this strategy just yet.

In brief, the model settings described in section 3 include a typical common good and a community of rational, asocial and memoryless agents maximizing lifetime utility by consuming the good. These agents are distinct from agents in most economic models insofar as they have a satiating consumption point with respect to the common good. Additionally, they are heterogeneous on many aspects: discount factors, satiation level and overall preferences. Of course, these novelties do not prevent them from falling in the trap of the tragedy of the commons. The baseline result for the community is a Nash equilibrium in which all agents consume their satiating consumption level, even though they would smooth consumption over time had they not strategic incentives to overconsume.

The story becomes more interesting when I give the agent the possibility to create a self-governance mechanism by paying the associated 1-period cost. The self-governance mechanism is a new agent called "Arbiter" that sets more sustainable consumption targets to move the community away from their current equilibrium path and potentially avert the tragedy of the commons. The Arbiter is characterized by an objective function maximized when there is no deviation between consumptions and the consumption targets and by a power budget. The Arbiter spends its power budget to dissuade the agents from consuming more than their consumption targets. It therefore alters the incentives of the agents in this community. I have also detailed the conditions for creating this Arbiter and emphasized how it requires a community heterogeneous enough to overcome the potential volunteer's dilemma.

The thesis has also evaluated the model against empirical facts. The model has made some inroads in providing a theory consistent with the evidence gathered by Ostrom (2000). This theory indeed predicts that the Arbiter is more likely to exist if it sets constraining consumption targets that take the level of the common goods into account, and if its punishments are proportional to derivations from targets rather than Arbitrary. More fundamentally, the model integrates the possibility of an institutional entrepreneur, an important precondition for reforms highlighted by Ostrom. The model however still requires significant improvement. Specifically, it contradicts conditional reciprocation by assumption. The model proposed here has no feedback mechanism between the agents and the Arbiter: it therefore fails to attribute fair targets and is not influenced by the community's preferences. This highlights a clear direction for future research: integrating a legitimacy-based budget to the model.

The theoretical model is novel in many ways. The agents are heterogeneous and have satiating consumption. They can also alter the institutional settings to self-constrain themselves for averting the tragedy of the commons. More fundamentally, with the treatment of a cooperation problem and the occurrence of a volunteer's dilemma, the model relies on a subtle combination of the game theoretical and structural approaches to move this line of research forward.
To the extent that the various assumptions are reasonable (empirical validity) and the conditions are respected (internal validity), it therefore provides novel insights on the issue of collection action surrounding a common good. I summarize these here.

With regards to solving the tragedy of the commons in general, the first and most important result is that rational, asocial, time-discounting agents can solve the tragedy of the commons under the circumstances detailed in the model, without relying on learning new norms or incomplete information. To the extent that these circumstances can be influenced by individuals, this brings hope that sustainable cooperation on collective action problems can be achieved by developing the proper incentive schemes, even with self-interested human nature. Fortunately, universal brotherly love or human-nature harmony are therefore not prerequisites. From an academic perspective, this also gives hopes to reconcile Economics with realistically more optimistic theoretical predictions.

Second, the creation of a self-governance mechanism crucially depends on a comparison between a single period's utility level with a lifetime utility differential between the scenario with and the one without self-governance. This implies that instantaneous or periodical comfort can be used as leverage to alter historical paths. This of course should be qualified when allowing for sustenance costs for the Arbiter, but overall, the transformation cost (of breaking the norms, bringing people together, …) is temporary, while the benefits persist for a community's lifetime. Concretely, one can hope to tilt the equilibrium path simply by providing a subsidy equivalent to a few periods of individual satiating consumption to a carefully selected subset of agents. This sensitivity of a community's institutions and history to one of its agents' temporary comfort is a double-edged sword. On one hand, it could allow well-meaning activists to solve collective action problems. On the other hand, it also empowers entrepreneurial agents to set up institutions that advantage them at the expense of the rest of the community. This could be studied more if a feedback mechanism was integrated in the model.

Third, the creation of a self-governance mechanism is delayed until the utility from future overconsumption and resource depletion is smaller than the utility from constrained consumption and more sustainable resource management. Timing is therefore key for institutional entrepreneurs: while acting too late obviously is suboptimal, acting too early is also likely to upset agents in the community and reduce aggregate lifetime welfare. While the model is very far from predicting the optimal date for holding a global Climate Change Response summit, its indications about timing were previously unseen in the CPR literature.

Finally, several specific results that appear robust can be recalled here. Ceteris paribus, an agent with a greater future orientation is more likely to create an Arbiter addressing the tragedy of the commons. Ceteris paribus, an agent whose Arbiter-set consumption target is not constraining her is more likely to create the Arbiter. Ceteris paribus, a community is more likely to create an Arbiter that has a bigger enforcement budget. Ceteris paribus, a community starting with a stock that is neither too big nor too small is more likely to create an Arbiter. I however expect some of these specific results to be qualitatively different if one gives the Arbiter a legitimacy-based budget.

These results and this exploratory model must be refined. However, it is a first theory that could help predict success of institutional entrepreneurship. By moving away from qualitative argument and forcing the theory onto a structural model, I hope this helps to better understand the complex phenomenon of self-organization for cooperation. I hope these early results help identify what determines whether one's activism efforts result in a dozen signatures on a petition in the bottom drawer of your desk or in a transnational agency recognized as expert authority by everybody, from the largest and least pacifist nations to rural households. But before making clear policy recommendations, a lot of work remains.

My intuition is that further research could be most fruitful by focusing on integrating a feedback mechanism from the community to the Arbiter. As discussed in section 5, this might be a challenge, but considering the applicability of the model to several prominent global threats and a multitude of local collective action problems, the rewards are well worth the effort. Overall, the model also suffers in terms of tractability from being set in discrete time. I regret not having had the foresight to build it in continuous time, as it would have allowed a more thorough analysis of community-specific results.

Re-orienting the research towards continuous time models would be helpful to future analysis of collective action problems.

More fundamentally, in line with previous findings, the role of institutions here is to alter incentives (Keohane & Ostrom, 1995). I model the new institution as exerting direct influence over agents' utility. However, the literature also emphasizes the role of institutions as information providers. Relaxing the assumption of full information in this model could therefore help explore circumstances under which information is useful to shift the path of the community (Keohane & Ostrom, 1995, Introduction).

Research relaxing the assumptions about the Arbiter –notably the exogenous nature of the Arbiter design- could allow to explore different situations. I have mentioned in the text the applicability of the model to political institutions and the Social Contract. Additionally, it could provide some insights in Competition Economics' political economy arguments. Competitive markets could indeed be considered as a tragedy of the commons under anarchy and creating a regulator could be perceived as a solution, like American broadcasting corporations lobbying for the creation of a Federal Radio Commission in the 1920s-30s.

One last important question I would like to address concerns empirical testing. To be of any use, the present model must be tested and refined with valid datasets beyond the Ostrom's facts. Institutions are difficult to study, especially when it comes to the international community where observations are scarce. This is another area where there is great potential for interdisciplinarity. Scholars of Sociology, Politics and International Relations have accumulated a wealth of relevant data over time. Applying econometric methods to these datasets could provide a good starting point for refining the model.

This paper is merely one stepping stone on the long road to understand collective action issues and develop solutions to these. Looking forward, after this model is structurally fine-tuned and finally validated, the obvious next step will be to integrate the determinants of the transformation cost. I suspect the identification of these determinants to rely mostly on Sociology and, ultimately, Psychology. After all, human brains are still the root decision-makers in any social system. But before our path finally reaches this area and connects to the intricacies of the mind, there are many more slippery stepping stones ahead that must be cleaned from the moss of ignorance.

Acknowledgments:

I am indebted to my supervisor Professor Duncan Snidal for his guidance and feedback – most of the IR and Politics components of this thesis are from his suggestions. I would like to thank Professor Godfrey Keller for his help with dynamic programming methods. I would also like to acknowledge you, dear reader, for going through over 100 pages of odd micro theory and still being so keen to even read the acknowledgments. Finally, this work would not have been possible without the encouragements of my family, my girlfriend and the irrational agents with prosocial preferences making up the anarchic community that is Linacre.

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2005). Institutions as a fundamental cause of long-run growth. *Handbook of economic growth*, *1*, 385-472.

Acheson, J. M. (2003). *Capturing the commons: devising institutions to manage the Maine lobster industry*. Hanover and London: University Press of New England.

Bardhan, P. (1999). Water community: An empirical analysis of cooperation on irrigation in south India. *Berkeley: University of California, Department of Economics, Working paper.*

Belleflamme, P. (2000). Stable Coalition Structures with Open Membership and Asymmetric Firms. *Games and Economic Behavior, 30,* 1-21

Bowles, S. (2004) *Microeconomics: Behavior, Institutions, and Evolution*. New Jersey, NJ: Princeton University Press.

Bueno De Mesquita, B., Smith, A., Siverson, R. M., & Morrow, J. D. (2005). *The logic of political survival.* MIT press.

Castillo, D., & Saysel, A. K. (2005). Simulation of common pool resource field experiments: a behavioral model of collective action. *Ecological Economics*, *55*(3), 420-436.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, *14*(1), 47-83.

Cox, M. (2014). Applying a social-ecological system framework to the study of the Taos Valley irrigation system. *Human ecology*, *42*(2), 311-324.

Crawford, E.S., and Ostrom, E. (1995). A grammar of institutions. *American Political Science Review, 89*(3) 582-600.

Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution*, *29*(4), 605-610.

Diekmann, A. (1993). Cooperation in an asymmetric volunteer's dilemma game theory and experimental evidence. *International Journal of Game Theory*, *22*(1), 75-85.

Dogan, M. (2002). Conceptions of legitimacy. In M. Hawkesworth & M. Kogan (Eds), *Encyclopedia of government and politics*, *1*, (pp. 116-126). London, Greater London: Routledge.

Editorial (26 October 2014) "The Guardian View on the Asian Infrastructure Bank: the US should work with it, not oppose it." *The Guardian*. Retrieved from https://www.theguardian.com

Feldhaus, C., & Stauf, J. (2016). More than words: the effects of cheap talk in a volunteer's dilemma. *Experimental Economics*, *19*(2), 342-359.

Gilmour, P. W., Dwyer, P. D., & Day, R. W. (2013). Enhancing the agency of fishers: a conceptual model of self-management in Australian abalone fisheries. *Marine Policy*, *37*, 165-175.

Hardin, G. (1968). The Tragedy of the Commons. *Science*, *162*(3859), 1243-1248.

Kaldor, N. (1957). A model of economic growth. *The Economic Journal*, *67*(268), 591-624.

Keohane, R. O., & Ostrom, E. (1995). *Local commons and global interdependence: Heterogeneity and Cooperation in Two Domains*. Sage Publications Ltd.

Lam, W. F. (1998). *Governing irrigation systems in Nepal: institutions, infrastructure, and collective action*. Institute for Contemporary Studies.

Ledyard, O. (1995). Public goods: some experimental results. In J. Kagel & A. Roth (Eds.), *Handbook of experimental economics*. Princeton: Princeton University Press (Chap. 2).

Libecap, G. D. (1995). The Conditions for Successful Collective Action. In *Local Commons and Global Interdependence: Heterogeneity and Cooperation in Two Domains*, ed. Robert Keohane and Elinor Ostrom, 161–90.

Morrow, C. E., & Hull, R. W. (1996). Donor-initiated common pool resource institutions: the case of the Yanesha forestry cooperative. *World Development*, *24*(10), 1641-1657.

Morrow, J. D., Bueno De Mesquita, B., Siverson, R. M., & Smith, A. (2008). Retesting selectorate theory: Separating the effects of W from other elements of democracy. *American Political Science Review*, *102*(03), 393-400.

Myatt, D. P., & Wallace, C. (2008). An evolutionary analysis of the volunteer's dilemma. *Games and Economic Behavior*, *62*(1), 67-76.

North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge university press.

Olson, M. (1965). *The logic of collective action*. Cambridge. Mass.: Harvard, 1971.
Ostrom, E. (1965). *Public entrepreneurship: a case study in ground water basin management* Doctoral dissertation, University of California, Los Angeles.

Ostrom, E. (2000). Collective Action and the Evolution of Social Norms. *Journal of Economic Perspectives*, *14*(3), 137-158.

Ostrom, E. (2009). A Polycentric Approach for Coping with Climate Change. *World Bank Policy Research Working Paper Series*, No. 5095

Ostrom, E. (2011), Background on the Institutional Analysis and Development Framework. *Policy Studies Journal, 39*: 7–27.

Pacheco, D. F., York, J. G., Dean, T. J., & Sarasvathy, S. D. (2010). The coevolution of institutional entrepreneurship: A tale of two theories. *Journal of management*, *36*(4), 974-1010.

Pierson, P. (2000). The limits of design: Explaining institutional origins and change. *Governance*, *13*(4), 475-499.

Poteete, A. R., Janssen, M. A., & Ostrom, E. (2010). *Working together: collective action, the commons, and multiple methods in practice*. Princeton University Press.

Przepiorka, W., & Diekmann, A. (2013). Individual heterogeneity and costly punishment: a volunteer's dilemma. *Proceedings of the Royal Society of London B: Biological Sciences*, *280*(1759), 20130247.

Ray, D., and Vohra, R. (2015). Chapter 5 – Coalition Formation. In Young, P., and Zamir, S. (Eds), *Handbook of Game Theory with Economic Applications, 4*. Elsevier.

Rockenbach, B., & Wolff, I. (2016). Designing Institutions for Social Dilemmas. *German Economic Review, 17*(3), 316-336.

Rustagi, D., Engel, S., & Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science*, *330*, 961-965.

Scharpf, F. W. (1997). Economic integration, democracy and the welfare state. *Journal of European public policy*, *4*(1), 18-36.

Viegas, F. B., Wattenberg, M., Kriss, J., & Van Ham, F. (2007). Talk before you type: Coordination in Wikipedia. In *System sciences, 2007. HICSS 2007. 40th annual Hawaii international conference on* (pp. 78-87). IEEE.

Weesie, J. (1994). Incomplete information and timing in the volunteer's dilemma: A comparison of four models. *Journal of Conflict Resolution*, *38*(3), 557-585.

Yi, S. S. (1997). Stable coalition structures with externalities. *Games and economic behavior*, *20*(2), 201-237.

Yi, S.S., and Shin, H.Z. (1995). "Endogenous Formation of Coalitions in Oligopoly," Dartmouth College, Department of Economics, WP No. 95-2.

Young, P. (2011). The dynamics of social innovation. *Proceedings of the National Academy of Sciences*, *108*(Supplement 4), 21285-21291.

Young, P. (2015). The evolution of social norms. *Annual Review of Economics*, *7*(1), 359-387.

Appendix

A-1.1 Solution for t^{*} and t^{**}

The transition rule of the stock is given by

$$S_{t+1} = (1+r) * \left(S_t - \sum\nolimits_{j}^{n} s_{t,j}\right) \tag{39}$$

I have argued in the text that agents will all consume their satiating consumption $\check{s}_i$. The transition rule therefore becomes

$$S_{t+1} = (1+r) * \left(S_t - \sum\nolimits_{j}^{n} \check{s}_j\right) \tag{40}$$

Since $\check{s}_i$ is constant over time and since I don't allow n to vary over time, for notational ease I set $\sum_j^n \check{s}_j = X$. The satiating consumption levels are always such that $0 < X$. In the case where $X \geq S_0$, the good is obviously exhausted at $t^* = 0$. I am therefore interested in the cases where $X < S_0$.

To determine t^*, I first convert (32) in its time invariant form by writing S_t as a function of S_0:

$$S_t = S_0(1+r)^t - X\sum_{z=1}^{t}(1+r)^z \tag{41}$$

I seek a value for t^*, which is the smallest integer such that $X \geq S_{t^*}$. I divide the analysis in cases conditional on r. As a starting point for each case, I use the condition that $X \geq S_{t^*}$, which implies:

$$S_0(1+r)^{t^*} - X\sum_{z=1}^{t^*}(1+r)^z \leq X \tag{42}$$

<u>Case 1:</u> $r = 0$

If $r = 0$, I have:

$$S_0 - X\sum_{z=1}^{t}(1)^z \leq X \tag{43}$$

$$S_0 - Xt \leq X \tag{44}$$

Since $X > 0$, this implies:

$$\frac{S_0 - X}{X} \leq t \tag{45}$$

This is a lower limit on t, which gives me a unique value for t^* since I am looking for the smallest t i.e. the earliest period satisfying that condition.

Case 2: $r > 0$

If $r > 0$, I replace the geometric sequence by its compact form, re-organize and obtain:

$$S_0(1+r)^{t^*} - X\frac{(1+r)^{t^*+1} - (1+r)}{r} \leq X \tag{46}$$

$$S_0(1+r)^{t^*} \leq \frac{X}{r}[(1+r)(1+r)^{t^*} - 1] \tag{47}$$

Dividing across by $(1+r)^{t^*}$ and re-organizing, I obtain:

$$(1+r)^{t^*}\left[S_0 - (1+r)\frac{X}{r}\right] < -\frac{X}{r} \tag{48}$$

The situation requires a further subdivision in two cases depending on the value of $S_0 - (1+r)\frac{X}{r}$. The special case where this term is equal to zero implies $r =$

$X/(S_0 - X)$ i.e. replenishment is equal to consumption rate. It leads to an indeterminacy: there is no t^*, which means the stock is never exhausted.

Case 2.1: $S_0 - (1+r)\frac{X}{r} > 0$

If $S_0 - (1+r)\frac{X}{r} > 0$ with $r > 0$ it implies $r > X/(S_0 - X)$, which places a lower limit on r. (48) therefore becomes

$$(1+r)^{t^*} < -\frac{X}{rS_0 - (1+r)X} \tag{49}$$

To solve for t^*, I would have to take the log of both sides of the inequality, which is possible only if the right-hand side's denominator is negative. This is the case only when $r < X/(S_0 - X)$, which is incompatible with the case we are in. There is therefore no t^* under these conditions. Specifically, when the growth rate of the good is too high compared to the aggregate satiating consumption, the good is never exhausted and the agents can fully enjoy the good.

Case 2.2: $S_0 - (1+r)\frac{X}{r} < 0$

With $r > 0$, this implies $r < X/(S_0 - X)$, which places an upper limit on r. Expression (48) becomes:

$$(1+r)^t > -\frac{X}{rS_0 - (1+r)X} \tag{50}$$

As for the previous case, taking log on both sides requires $r < X/(S_0 - X)$, which is fine in this case. Given that $r > 0$, $ln(1+r)$ is positive, so that I obtain a lower limit on t:

$$t > \frac{ln(X) - ln\,(X - r(S_0 - X))}{ln\,(1+r)} \tag{51}$$

The lowest integer t that satisfies this inequality is the t^* of the model under these conditions.

Case 3: $r < 0$

The fact that $r < 0$ does not affect the manipulations bringing me to (48). I therefore start with

$$(1+r)^t\left[S_0-(1+r)\frac{X}{r}\right]<-\frac{X}{r} \tag{52}$$

As previously, I subdivide further into 2 cases.

Case 3.1: $S_0-(1+r)\frac{X}{r}>0$

If $S_0-(1+r)\frac{X}{r}>0$ with $r<0$, I have $r<X/(S_0-X)$, which places an upper limit on r. (52) therefore becomes

$$(1+r)^t<-\frac{X}{rS_0-(1+r)X} \tag{53}$$

Taking log on both sides requires $rS_0-(1+r)X<0$ which implies $r<X/(S_0-X)$. Since $r<0$, this is always satisfied. In order to have a real value for the log I need $(1+r)>0$, which implies that $-1<r$. This makes intuitive sense since $r=-1$ would imply that the good vanishes in period 0, so that $t^*=0$. With $-1<r<0$, I have $ln(1+r)<0$, so that I obtain the following lower limit on t:

$$t>\frac{ln(X)-ln\,(X-r(S_0-X))}{ln\,(1+r)} \tag{54}$$

The lowest integer t that satisfies this inequality is t^* under these conditions.

Case 3.2: $S_0 - (1+r)\frac{X}{r} < 0$

With $r < 0$, $S_0 - (1+r)\frac{X}{r} < 0$ would imply $r > X/(S_0 - X)$. Given our conditions on X and S_0 i.e. since $0 < X < S_0$, this case is impossible, in the sense that it cannot occur and does not need to be evaluated.

Summary for t^*:

I have therefore established the value of t^* in 5 different cases. Because it can only take integer values,

$$t^* = roundup(\varepsilon^*) \tag{55}$$

With the following values for ε^*:

$\boldsymbol{r}$	→	$\boldsymbol{\varepsilon}^*$
$r = -1$	→	0
$-1 < r < 0$	→	$\frac{\ln(X) - \ln(X - r(S_0 - X))}{\ln(1 + r)}$
$r = 0$	→	$\frac{S_0 - X}{X}$
$0 < r < \frac{X}{(S_0 - X)}$	→	$\frac{\ln(X) - \ln(X - r(S_0 - X))}{\ln(1 + r)}$
$\frac{X}{S_0 - X} \leq r$	→	$+\infty$

The procedure for t^{**} is the same, except that $\sum_j^n(\check{s}_j - p_j) = X$ and I must replace S_0 by S_x with $S_x = S_0(1 + r)^x - \sum_j^n \check{s}_j \sum_{z=1}^x(1 + r)^z$

A-1.2 Solution to the Arbiter's problem

As explained in the text, the Arbiter faces the following problem every period:

$$\max_{p_{t,i}\forall i\in[1,\ldots,n]} \mathrm{L}(s'_{t,i}) = -\sum_i^n \left(s^*_{t,i} - s'_{t,i}\right)^2$$
$$subject\ to$$
$$\mu\sum_i^n p_{t,i} \leq P \tag{56}$$
$$s'_{t,i} = \hat{s}_{t,i} - p_{t,i} \qquad for\ all\ i$$
$$0 \leq p_{t,i} \leq \hat{s}_{t,i} \qquad for\ all\ i$$

The constraints on $p_{t,i}$ form a compact set, so by Bolzano-Weierstrass thereom, there must exist a global maximum. The constraint qualification cannot fail so long as $\frac{P_t}{\mu} > 0$ and $0 \neq \hat{s}_{t,i}$ hold for all i. Note that if either of these conditions did not hold, the Arbiter's would not be able to choose any $p_{t,i} \neq 0$ anyway. The Kuhn-Tucker conditions are therefore necessary to find the maximum. I set up the following Lagrangean:

$$\begin{gathered}\mathcal{L} = \\ -\sum\nolimits_{i}^{n}\left(s_{t,i}^{*} - s'_{t,i}\right)^{2} - \lambda\left(\sum\nolimits_{i}^{n} p_{t,i} - \frac{P}{\mu}\right) + \sum\nolimits_{i}^{n} \gamma_i p_{t,i} \\ -\sum\nolimits_{i}^{n} \varphi_i\left(p_{t,i} - \hat{s}_{t,i}\right)\end{gathered} \tag{57}$$

Where $\lambda, \gamma_1, \dots, \gamma_n, \varphi_1, \dots., \varphi_n$ are the Lagrangean multipliers. Plugging in $s'_{t,i} = \hat{s}_{t,i} - p_{t,i}$ and taking the first order conditions gives:

$$\begin{gathered}-2p_{t,i} + 2\hat{s}_{t,i} - 2s_{t,i}^{*} - \lambda + \gamma_i - \varphi_i = 0 \quad for\ all\ i \\ \in [1, 2, \dots, n]\end{gathered} \tag{58}$$

The complementary slackness conditions are:

$$\lambda^* \geq 0; \sum\nolimits_i^n p_{t,i}^* \leq \frac{P}{\mu} \; and \;\; \lambda^* \left(\sum\nolimits_i^n p_{t,i}^* - \frac{P}{\mu}\right) = 0$$

$$\gamma_i^* \geq 0; -p_{t,i}^* \leq 0 \; and \;\; \gamma_i^*\left(-p_{t,i}^*\right) = 0 \; for \; all \; i \in [1,2,\dots,n] \qquad (59)$$

$$\varphi_i^* \geq 0; p_{t,i}^* \leq \hat{s}_{t,i} \; and \;\; \varphi_i^*\left(p_{t,i}^* - \hat{s}_{t,i}\right) = 0 \; for \; all \; i \in [1,2,\dots,n]$$

I decompose the analysis in several cases. First, I distinguish between the cases where the budget constraint binds. Within both subsets of cases, there is a range of cases where the optimal sanction is 0 for 1, 2, … n-1 or n agents. This determines how many γ_i^* are 0 in equilibrium. To encompass the entire range of cases at once, I assume my indexing $i \in [1,2,\dots,n]$ is ordered with those agents for whom $\gamma_i^* = 0$ indexed by the lowest numbers. I denote them by c = 1, …, C. I therefore have 0 ≤ C ≤ n. For C equations, the optimal sanction is therefore positive, so that $\gamma_c^* = 0$ for all $c \in [1,2,\dots,C]$. It is straightforward that the n-C others face sanctions $p_{t,n-c}^* = 0$ by the complementary slackness condition. For these, the upper limit $p_{t,i}^* \leq \hat{s}_{t,i}$ never binds, so that there is no further subdivision according to whether or not φ_i^* is zero. However, this multiplier matters for the those receiving a positive punishment, as I show in each case.

Case 1: the budget constraint binds

Given that $\frac{P}{\mu} > 0$, a binding budget constraint implies that C > 0. Indeed, at least one of the sanctions must be positive to exhaust a positive budget. Furthermore, among these positive $p_{t,c}$, I order the agents such that the first A of them are bound by the constraint $p_{t,c} \leq \hat{s}_{t,c}$, and therefore decide $p_{t,c} = \hat{s}_{t,c}$. With the complementary slackness conditions, I therefore obtain the following system of equations:

$$\begin{array}{ll} -2p_{t,c} + 2\hat{s}_{t,c} - 2s^*_{t,c} - \lambda = 0 & for\ all\ c \in [A+1, \dots, C] \\ 0 < p_{t,c} & for\ all\ c \in [1, \dots, C] \\ p_{t,c} < \hat{s}_{t,c} & for\ all\ c \in [A+1, \dots, C] \\ p_{t,c} = \hat{s}_{t,c} & for\ all\ c \in [1, \dots, A] \\ p_{t,n-c} = 0 & for\ all\ n-c \in [C+1, \dots, n] \\ \multicolumn{2}{c}{p_{t,c} = \frac{P}{\mu} - \sum_{j=1, j \neq c}^{C} p_{t,j}} \end{array} \tag{60}$$

Notice that the A agents for whom $p_{t,c} = \hat{s}_{t,c}$ are just a special case of positive sanction where the Arbiter's target is set to 0. Effectively, the conditions on the targets of the Arbiter (especially $0 \leq s^*_{t,i}$) implies this constraint is always satisfied. I can therefore consider the A agents the same way I consider the other agents with positive punishments.

This problem can then be solved with a bit of algebra to obtain the answer reported in the text:

$$p_{t,c}^{*} = \frac{1}{C}\left[\frac{P}{\mu} + (C-1)\left(\hat{s}_{t,c} - s_{t,c}^{*}\right) - \sum\nolimits_{j=1,j\neq c}^{C}\left(\hat{s}_{t,j} - s_{t,j}^{*}\right)\right]$$

$$for\ all\ c\ \in [1,2,\dots,C]$$

(61)

$$and$$

$$p_{t,n-c}^{*} = 0\ \ for\ all\ n-c \in [C+1,\dots,n]$$

As reported in the text.

Case 2: the budget constraint does not bind

Contrary to case 1, this case is consistent with 0 ≤ C ≤ n. As in the case where the budget constraint binds, conditions on the Arbiter's targets allow me to analyze agents with positive punishments at once. With the complementary slackness conditions, I again obtain 2 sets of rewritten FOCs:

$$\begin{aligned} -2p_{t,c} + 2\hat{s}_{t,c} - 2s_{t,c}^{*} &= 0 \quad for\ all\ c\ \in [1,2,\dots,C] \\ p_{t,n-c} &= 0 \qquad\qquad for\ all\ n-c \in [C+1,\dots,n] \end{aligned} \tag{62}$$

It is straightforward to show that in that case:

$$p^*_{t,c} = \hat{s}_{t,c} - s^*_{t,c} \quad for\ all\ c\ \in [1, 2, \dots, C]$$
$$p^*_{t,n-c} = 0 \quad for\ all\ n - c \in [C + 1, \dots, n] \tag{63}$$

These 2 cases cover the whole set of possibilities assuming $\frac{P_t}{\mu} > 0$. Here, $\hat{s}_{t,c} = s^*_{t,c} \rightarrow p_{t,c} = 0$. The n-c special cases are encompassed by the more general formula for $p^*_{t,c}$ and therefore do not have to be considered separately.

What determines whether the Arbiter's given budget constraint binds is the sum of deviation from target in the community $\sum_i^n(\hat{s}_{t,i} - s^*_{t,i})$. The budget binds when $\sum_i^n(\hat{s}_{t,i} - s^*_{t,i}) > {}^P\!/_\mu$.

A-1.3 Effect of satiating consumption level on CCA:

Using the fact that the derivative of an injective function is equal to the reciprocal of the derivative of its inverse, I can express the effect of $\check{s}_i$ on the right-hand side (RHS) of the CCA$_i$ as follows:

$$
\frac{\delta \mathrm{RHS}}{\delta \check{s}_i}
= 1
- \frac{u_i'(\check{s}_i)\sum_{t=x}^{t^*-1}\beta_i^{t-x} - u_i'\left(\check{s}_i - \frac{1}{n}\left[\frac{P}{\mu} + (n-1)(\check{s}_i - s_i^*) - \sum_{j=1,j\neq i}^{n}\left(\check{s}_j - s_j^*\right)\right]\right)\frac{1}{n}\sum_{t=x+1}^{t^{**}-1}\beta_i^{t-x}}{u_i'(Z)}
$$

Where

$$
\begin{aligned}
Z = \sum_{t=x}^{t^*-1}\beta_i^{t-x}u_i(\check{s}_i) + \beta_i^{t^*-x}u_i\left(\hat{s}_{t^*,i}\right) + \sum_{t=t^*+1}^{t^{**}}\beta_i^{t-x}u_i(0) \\
- \sum_{t=x+1}^{t^{**}-1}\beta_i^{t-x}u_i\Bigg(\check{s}_i \\
- \frac{1}{n}\left[\frac{P}{\mu} + (n-1)\left(\check{s}_i - s_{t,i}^*\right) - \sum\nolimits_{j=1,j\neq i}^{n}\left(\check{s}_j - s_{t,j}^*\right)\right]\Bigg) \\
- \beta_i^{t^{**}-x}u_i\left(\hat{s}_{t^{**},i}\right)
\end{aligned}
\tag{64}
$$

Effectively, the effect of a marginally higher $\check{s}_i$ is ambiguous. If it is positive, it will make it more likely for $\mathrm{CCA_i}$ to hold. This is only valid within a single-time step, so a numerical analysis might be more appropriate. Using $u_i(s_i) = \ln(1 + s_i)$ and a calibration with reasonable values, I obtain the following graph for $RHS(\check{s}_i)$:

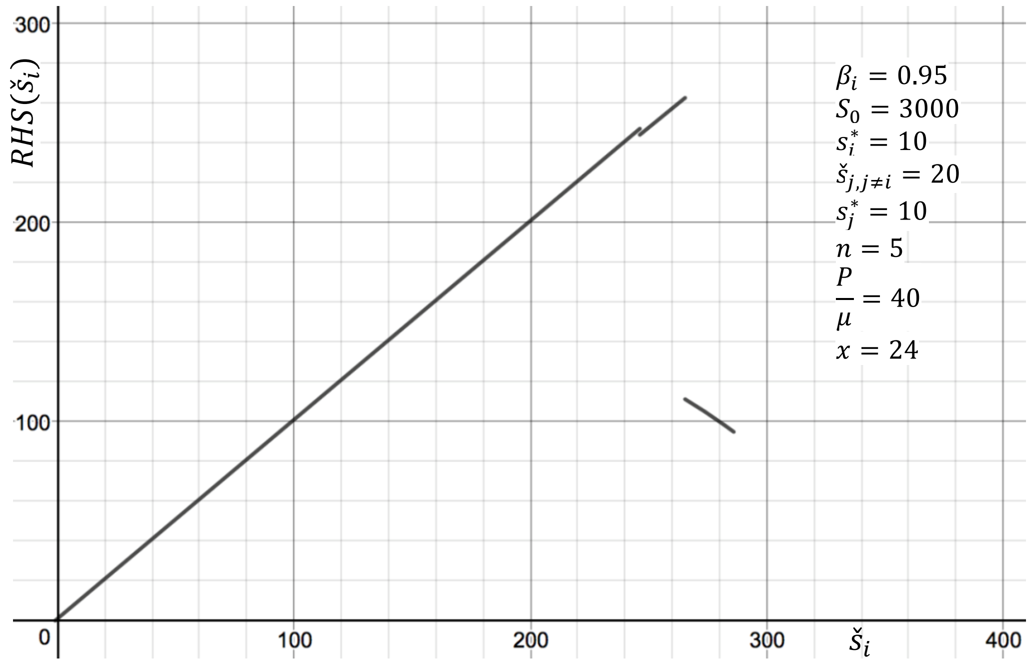

Figure A.1: Effect of agent i's satiating consumption on her propensity to create the Arbiter, with $u_i(s_i) = ln\,(1 + s_i)$

However, there are significant interaction effects: the shape of the graph is qualitatively different under different specifications (with the same utility function and still satisfying the conditions for a determinate equilibrium). In both figures A.2 and A.3 below, the condition for a determinate equilibrium holds. Yet, we see that whether a marginal increase in satiating consumption increases the likelihood of CCA to hold depends on the current level of satiating consumption.

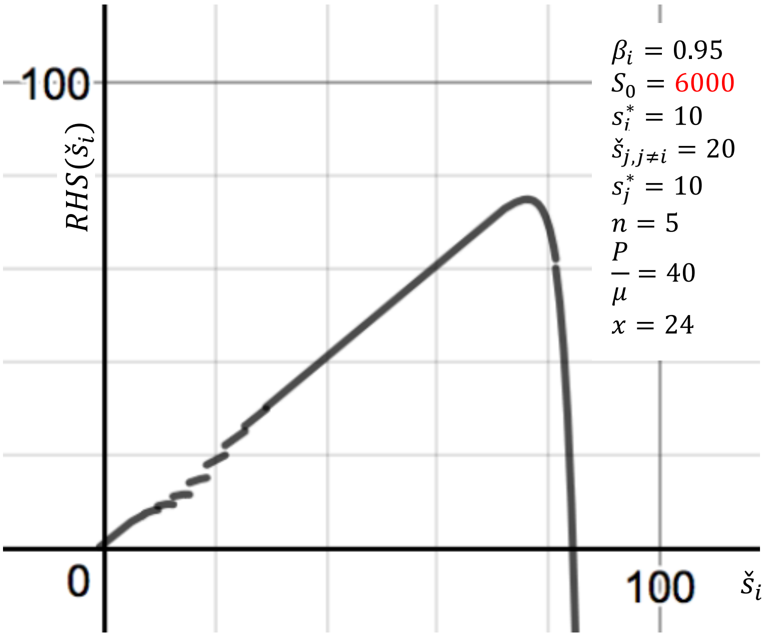

Figure A.2: Sensitivity of results to greater initial stock of good

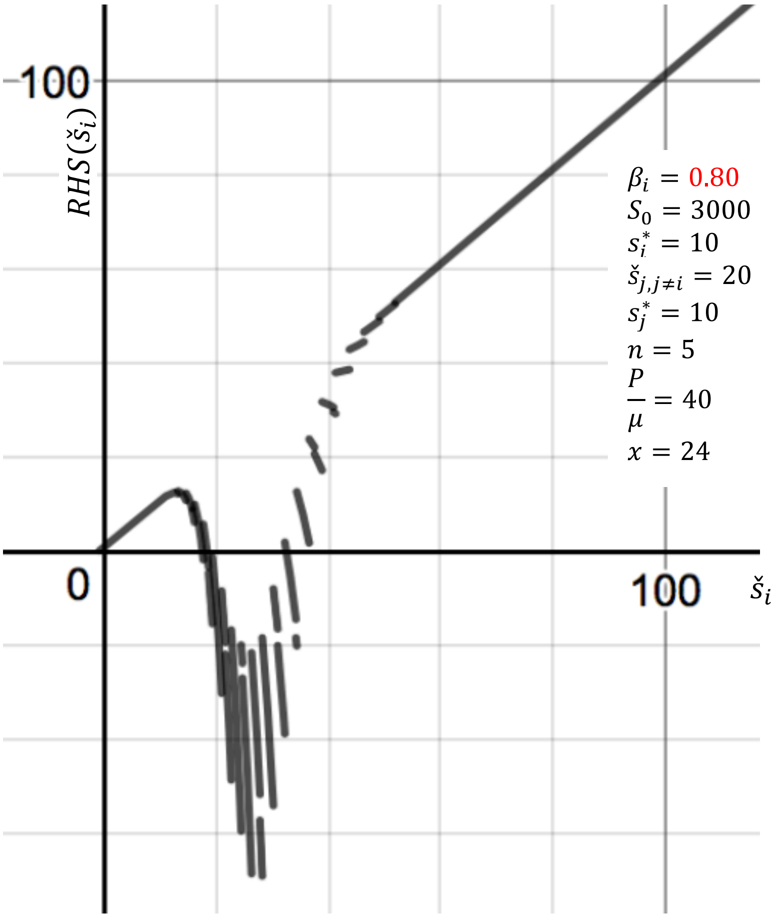

Figure A.3: Sensitivity of results to smaller discount factor