Morality and Equality from Rationality Alone - A repeated game approach of contractarianism

Alexis Louaas

Morality and Equality from Rationality Alone
-
A repeated game approach of contractarianism

Alexis Louaas*

September 24, 2020

Abstract: This paper highlights the role that equality and reciprocity play in preserving peace and cooperation among individuals with conflicting interests. Following the contractarian tradition, I model a mutually beneficial interaction as a prisoner's dilemma and using repeated game theory, I show that a mutually beneficial joint venture may be undertaken only if the final distribution of incomes is sufficiently egalitarian. From a pre-moral context, the model allows to derive endogenous bounds on the income of each individual that reproduce Moehler (2018)'s *weak universalisation principle*. Contrasting with the well-known equity-efficiency trade-off, the model also produces an equity-efficiency complementarity.

Keywords: contractarianism, morality, equity, efficiency, game theory, norms

*CREST-Ecole Polytechnique, France. E-mail: alexis.louaas@polytechnique.edu

1 Introduction

Can morality be a means of common progress and advancement? The contractarian theory, that seeks to derive moral norms from the sole assumption of instrumental rationality, allows to tackle this fundamental question. Although its lack of historical and cultural perspectives undermines its descriptive power, the contractarian approach has a strong normative appeal because it ultimately relies on the shared rationality of those who interact under a given set of moral norms. Norms that can rationally be justified may hence pretend to a large degree of objectivity, even when individuals hold different cultures and histories. This makes contractarian theory particularly relevant when individuals hold *a priori* diverse moral principles stemming from diverse cultural, socio-economic and historical backgrounds, as emphasised by Moehler (2020). But the test of rationality may also constitute a potent argument for positive change in a society of individuals with similar values, looking for new ways of interacting. Contractarian theory hence offers an interesting perspective on how moral norms may spur progress.

In order to achieve this, contractarians must nevertheless show that instrumental rationality can indeed be used to derive at least some of the moral principles that currently govern our societies. In the words of Thoma (2015), contractarians need to show that

> The rules that we commonly think of as constituting morality would be agreed-upon by rational and purely self-interested individuals in a pre-moral context.

The present paper contributes to this project by highlighting the role that equality and reciprocity play in preserving peace and cooperation among players with conflicting interests. I model two agents who jointly produce an output that they could not produce separately. The two agents bargain over the distribution of this output and define property rights. This bargaining process is exogenous and yields an outcome that is morally unconstrained, where each individual receives a share of the output that depends only on his relative bargaining powers. Once production has taken place, each agent may either respect the existing property rights or dispute them in view of obtaining a larger output share. Dispute by both players leads to Hobbes' war of all against all, where individuals waist large amounts of resources because of conflict. In contrast, the rule of law prevails if both agents respect the property rights previously decided. Both players therefore have an *overarching goal* to preserve peace and avoid the high cost of conflict even though the Nash equilibrium of the stage game is the war of all against all. However, I show that sufficiently *patient* agents are able to escape the war of all against

all when the game is repeated an indefinite number of time if, and only if, the distribution of the final output is not too unequal and if agents adopt reciprocal behaviours. Since both agents are better-off under the rule of law, it is in their own personal interests to constrain their bargaining power to avoid conflict. In particular, an agent with a high bargaining power will always find it better to limit his share of the total output to a level that guarantees cooperation from the other player. This condition formalises Moehler (2018)'s *weak universalisation principle* of *"each according to her basic needs and above this level according to her relative bargaining power"* and fits the justification provided in Moehler (2020) (p 50)

> If agents do not satisfy the principle [], then, considered from their own perspectives, the agents must expect that the bargaining process will break down and peaceful long-term cooperation is threatened. Failure of coordination in cases of conflict in the strict sense [] leads to a breakdown of cooperation and destructive action, and thus, from the perspective of *homo prudens*, to a worst outcome than compliance with the principle.

Gaus (2019) and Messina & Wiens (2020) have argued that previous contractarian approaches, including the seminal work of Gauthier (1986) and more recently Moehler (2018), fail to build morality on the basis of a truly pre-moral setting because they impose on the bargaining process (and therefore on the solution) constraints, such as the symmetry of the players' strategies, that have no empirical or theoretical reason to hold and should in fact, be considered as morally motivated. The symmetry of the outcome being the result of these morally motivated constraints, Gauthier (1986) and Moehler (2018)'s models cannot be credited with generating endogenously moral norms. In contrast, the moral constraints obtained in this paper do not rely on any assumption restricting the bargaining procedure but on non-moral assumptions related to the characteristics of the agents, such as their degree of patience, and to their joint production opportunities.

In addition to its affiliation to the contractarian literature the present paper bridges a gap with the literature that, following Axelrod (1981), seeks to derive norms from evolutionary game theory. By showing how, in large groups of individuals, some cooperative strategies consistently out-preform others, this literature explains the emergence of norms as dominant social practices. The present paper shows that too much inequality prevents the emergence of cooperation. Nevertheless, the fact that cooperation can happen does not mean that it has to happen.[1] Evolutionary game theory pro-

[1]The equilibrium concept used in this paper is Subgame Perfection, which notoriously

vides the missing theoretical argument for which cooperation is likely to emerge when it is made possible by an adequate setting (low inequality and high patience in this paper): the advantage that cooperation provides, in terms of wealth and satisfaction, facilitates its diffusion across society and allows cooperation to become a dominant norm.[2] While emphasising the adverse effect of inequality on cooperation, the present paper can therefore be seen as a building block for the wider theory of norm adoption. In particular, its results are compatible with the mutualistic approach defended in Baumard et al. (2013).

The insights derived hereafter are also related to the experimental literature on the determinants of cooperation, in which the experimenter is able to impose exogenous variations on the rules of games played by participants in order to identify the conditions most favourable to cooperation. The relationship between the current paper and this literature is bidirectional. On the one hand, Dal Bó & Fréchette (2018)'s meta-analysis of experimental prisoner's dilemmas shows a significant and negative relationship between the minimal patience level δ^*, from which cooperation may occur, and the actual frequency at which players indeed cooperate. This suggests that δ^* is indeed a natural metric for the probability that individuals will cooperate.[3] On the other hand, the outcomes of these experimental games are not mere results of the artificial conditions created by the experimenter, but also of the more general conceptions of fairness and justice that participants may hold. Understanding cooperation therefore requires comings and goings between general theories of moral norms, such as developed in this paper, and applied experiments.

Finally, the results presented below should be related to the wider debates on inequalities and on those fields of politics, philosophy and economics more specifically concerned with the relationship between equality and efficiency. Economic theory often assumes *a priori* that incomes reflect individual contributions to the joint output. Concerns for equality are often seen either as exogenous norms that constrain the proper functioning of the economy,[4] or

yields multiple outcomes. In particular, the Nash Equilibrium where agents fail to cooperate is always always a Subgame Perfect Equilibrium (SPE). Even though, it seems natural that, given the possibility of mutually beneficial cooperation, individuals would cooperate, the concept of SPE equilibrium does not rule out the possibility that they fail to do so.

[2]In other words, evolutionary game theory provides an equilibrium selection criteria that solves the multiplicity issue. Modelling explicitly the evolutionary dynamics of the game presented here is an interesting venue for further research but lies outside the scope of the current paper.

[3]From a theoretical standpoint, this does not resolve the multiplicity issue but it dampens its empirical relevance for the theory presented here.

[4]This is the case of the optimal taxation literature initiated by Mirrlees (1971). For a

as rent-seeking behaviours from poorer individuals seeking to appropriate a share of the output produced by richer individuals. In these two cases, more equality is associated with lower efficiency because high income individuals reduce their (highly productive) work effort in response to re-distributive policies that despoil them from their rightful income. The exact opposite is true in the present model. Individuals are assumed to contribute equally to the common good but they are allowed to bargain unequal shares of the total cooperative output. As a consequence, the poorer individual's contribution is higher than his income. When the wedge between his production and his income becomes too large, he stops cooperating, hence precipitating the war of all against all. This gives rise to an equality-efficiency *complementarity*. Contrasting with theories in which norms of equality are detrimental to economic efficiency, equality is here presented as a social technology that allows individuals to exploit the gains from joint effort. This equality-efficiency complementarity contradicts a mechanism at the heart a vast economic literature, currently puzzled by the historical and cross-sectional negative relationships observed between inequality and growth.[5]

The paper is organised as follows. Section 2 describes the stage game of the model. Section 3 introduces the indefinitely repeated version of the game and presents the minimal conditions under which cooperation emerges. Section 4 extends the model to a large number of agents to investigate the emergence of norms in a population and Section 5 characterises weaker conditions under which cooperation may take place. Finally, Section 6 concludes.

2 Stage game

Two agents can collectively produce an output $R > 0$, net from the costs of production (labour, resources, etc). Before starting the production process, they decide how to share the output R. The resulting property right system hence allocates a share ϕ to the column player while the row player receives the share $1-\phi$. No assumption is made on the bargaining procedure that delivers this distribution. For convenience, one may think of Nash Bargaining but none of the results below depends on the exogenous bargaining procedure considered. For any bargaining procedure, this paper is only concerned with finding bounds on the share of the total output that an individual will claim, no matter how high his bargaining power. This self-enforced restric-

more recent review of the literature see Piketty & Saez (2013)

[5]For a recent example see Berg et al. (2018). See also Aghion et al. (1999) for an authoritative although a little outdated review of both empirical and theoretical literatures on the relationship between inequality and growth.

tion comes, as will be shown below, from the fear of the war of all against all triggered by a defection of the poorer individual.

The game is the following. Once the bargaining procedure has taken place and a property right scheme $(\phi, 1-\phi)$ is set-up, agents may comply (C) with the agreed-upon property right system and receive their respective shares ϕ and $1-\phi$, or dispute (D) it. If an agent unilaterally disputes the scheme, he takes the full cooperative output R, leaving nothing to the other player. When both agents defect, they engage in a conflict that costs them each c and the surplus is shared equally among them. This symmetry reflects Hobbes' view of "rough natural equality among agents" (Moehler (2020): 12) in the war of all against all state of nature.[6][7]

The payoff of the stage game is represented in Table 1. The assumption $R/2 - c > 0$ guarantees that the conflict does not exhaust the benefits from cooperation. As a consequence, both agents prefer to fight rather than to let the other player unilaterally dispute the property right scheme, hence reaping the full cooperative output.

		C	D
2*	C	$\phi R, (1-\phi)R$	$0, R$
	D	$R, 0$	$R/2 - c, R/2 - c$

Table 1: Stage game, with $R/2 - c > 0$

The total output is always larger when individuals comply with the rule of law and therefore save themselves from costly conflict. However, the unique Nash Equilibrium of the game is the war of all against all, represented by the pair of strategies (D, D).

When

$$\frac{1}{2} - \frac{c}{R} \leq \phi \leq \frac{1}{2} + \frac{c}{R}, \tag{1}$$

[6]In some settings, the "rough natural equality among agents" may not be a good assumption. It is therefore relaxed, and its role precisely discussed in Section 3.1.

[7]The game presented in Table 1 is designed to fit Hobbes' theory of the social contract. In particular, it makes explicit reference to the war of all against all. However, Appendix 6.1 shows that the model can be re-written to accommodate an interpretation where the property rights cannot be questioned, but both players can shirk and free-ride on the other players' efforts. While further away from Hobbes' original account of the social contract, this alternative interpretation may appear slightly more attractive to readers accustomed problems framed in terms of incentives to work, rather than in terms of incentives to preserve order.

both agents are better-off if they comply with the property right scheme.[8] Contrasting with Moehler (2018)'s informal argument and Messina & Wiens (2020)'s formal representation of the Hobbesian framework, the game summarised in Table 1 is, in this case, a prisoner's dilemma. Individual rationality leads both agents to defect even though it would be profitable for both of them to cooperate (given that the other player cooperates), because none of them can be sure that the other will indeed cooperate. To solve this particularly frustrating issue, known as the assurance problem, Hobbes concluded in his "second law of nature", to the necessity for individuals to transfer their rights of nature to "do what they consider necessary for survival and transfer these rights to a common authority that is not part of the society" (Messina & Wiens (2020): 20). This conclusion can be supported in the present framework by assuming that an institution has the power to inflict a cost f to any defector. A sufficiently high cost $f \geq R \max(\phi, 1-\phi)$, would indeed discourage defection and coordinate individuals on the Pareto-superior equilibrium (C, C). This solution to the assurance problem has the merit of fitting nicely with the institutions that have governed our societies. Nevertheless, it fails to provide an account of how individuals would build such an institution and why they would respect it. The neutrality of this external authority should also be questioned: why would an authority sufficiently power-full to inflict a high cost on any defectors, not use his power to extort his subjects?

Avoiding these short-comings requires to either endogenize the external authority or to dispense with it altogether. A community of more than two agents has indeed the possibility to build formal institutions, external to potential conflicts, whose legitimacy relies on a consensus among a majority of agents, and that are in charge of sanctioning inadequate behaviours. In contrast, the two-players framework presented here, does not allow to sustain such external institutions. The norms of behaviours that appear must therefore be self-sustaining. The theory presented here is consequently much more a theory of norms than a theory of formal institutions, even though norms and formal institutions may, in practice, have the same purpose of maintaining cooperation.

[8] When the property right system is such that either $\phi > 1/2 + c/R$ or $\phi < 1/2 - c/R$, that is one agent receives a very large proportion of the cooperative surplus, there is no obvious way to rank (C, C) and (D, D). The total output remains higher under the rule of law, but the least well-off agent receives a share of the output that is so small that he prefers the war of all against all to the cooperative state. Individual rationality conducts both agents to defect but the cooperative state (C, C) is not Pareto superior in this case. The game can therefore not be called a prisoner's dilemma. We will see that, despite the very high inequality of bargaining powers, this case may nevertheless give rise to endogenous moral norms. See footnote 10.

3 Indefinitely repeated games

The two players indefinitely repeat the stage game summarized in Table 1. Future payoffs are discounted at a rate $\delta \in (0, 1)$ to reflect a preference for immediate rewards over future ones. An alternative interpretation is that, at each stage, the agents assign a probability δ that the game will continue one additional period. Under these two interpretations, δ represents the value that the agent gives to payoffs resulting from possible future interactions, and hence captures his degree of *patience*.

The concept of Subgame Perfect Equilibrium (SPE) will be used to characterise possible outcomes of the repeated game. Subgame perfection is a refinement of the Nash Equilibrium, in which all strategies adopted by the players are required to be Nash Equilibrium of each sub-game of the repeated game. This, in particular, rules out strategies that involve non credible threats, and is considered to be the most natural equilibrium concept in a repeated interaction framework where individuals are rational and self-interested.

Among the many strategies that can be supported as SPE, the grim trigger (or grim strategy) plays a particular role. Under a grim trigger, a player chooses to comply with the property right system on the first period, and then complies as long as the other player does. If a deviation is observed, the player retaliates by playing D for the rest of the game, resulting in a perpetual state of war of all against all. Playing D is a credible threat since it is a Nash equilibrium of the stage game (see for example Fudenberg & Tirole (1991)). Grim strategies are interesting focal points because they define minimal conditions under which cooperation is possible. Since they entail the hardest punishment possible (playing D for ever after a deviation is spotted), they require less patience than other strategies to be sustained as SPE. In addition, they entail symmetrical punishments which, as shown below implies that the condition on the minimal level of patience such that they are SPE, can also be expressed in terms of inequality.

3.1 Grim strategies

The pair of grim strategies is a SPE if and only if none of the players have a profitable deviation when they play C,[9] which may be written mathematically for the row player as

$$R - \phi R \leq \sum_{t=1}^{\infty} \delta^t (\phi R - (R/2 - c)).$$

[9]For a general and in-depth treatment of game theory, see Fudenberg & Tirole (1991).

That is, the short-term gain from deviating from C to D when the other player plays C, represented on the left-hand side of the inequality, has to be smaller than the foregone future gains from cooperation, represented by the right-hand side of the inequality. Replacing ϕ by $1-\phi$ in the equality gives the condition under which deviation is not a profitable strategy for the column player.

Equivalently, these two conditions define a lower bound on the discount factor

$$\delta \geq \max\Big(\frac{R(1-\phi)}{R/2+c}, \frac{R\phi}{R/2+c}\Big) \equiv \delta^*(\phi), \tag{2}$$

such that (C,C) is a SPE when $\delta \geq \delta^*(\phi)$. Cooperation emerges as a possible outcome of the interaction where each agent renounces to the short-term gains from unilateral deviation in order to improve *his own* long-term satisfaction.

The lower bound $\delta^*(\phi)$ reaches its minimal value when $\phi = 1/2$, that is, when the property right system attributes the same share of the cooperative output to the two players. More generally, the parameter space that supports (C,C) as a SPE expands as ϕ gets closer to $1/2$, i.e. when there is less inequality. Condition (2) can therefore also be seen as a constraint on the maximum inequality compatible with cooperation. Indeed, as the inequality of property right system increases, it becomes harder to sustain cooperation as a SPE for a given value of δ. The intuition beyond this result is simple: under a grim trigger scheme, mutual cooperation is a sustainable and rational behaviour only if the long-term gains from future cooperation are higher than the short term gain from unilateral deviation. While the short term gains from deviation are unaffected by the characteristics of the property rights system, the long-term gains from cooperation heavily depend on it. For a given individual, the more generous his share in the cooperative surplus, the lower his degree of patience must be to justify forgoing the short-term deviation gain. Since cooperation must be in the best interest of both players, it is the situation of the least well-off individual that determines whether mutual cooperation is a SPE or not. Consequently, if a player receives a higher share of the cooperative output than the other, decreasing this share lowers the patience threshold $\delta^*(\phi)$ above which cooperation is a SPE.

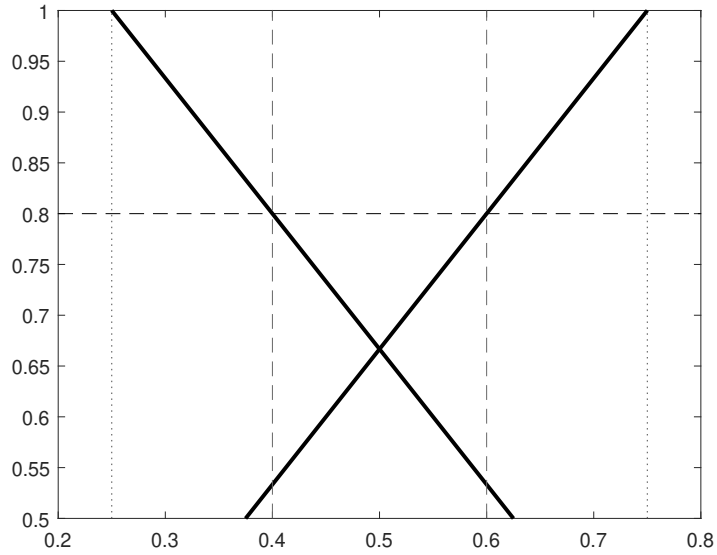


Figure 1: The values of $\phi \in [0,1]$ are read on the x-axis and the relevant values of δ are read on the y-axis. The highest of the two dark lines represents the function $\delta^*(\phi)$. The horizontal dashed line represents a possible value (0.8) of the rate of patience. It crosses the graph of $\delta^*(\phi)$ at the two points of abscissas $1-\delta\frac{R/2+c}{R} = 0.4$ and $\delta\frac{R/2+c}{R} = 0.6$, that define the largest inequality compatible with (C,C) being a SPE. The computations were realised with $R = 4$, $c = 1$ and $\delta = 0.8$.

The importance of equality for the emergence of cooperative behaviours can be better understood by taking a slightly different perspective. Instead of taking ϕ as fixed and let δ vary as previously, assume that the players have a given rate of patience $\delta \in [\delta^*(1/2), 1)$. Then, re-writing inequality (2) shows that (C,C) is a SPE if and only if

$$1 - \delta\frac{R/2+c}{R} \;\leq\; \phi \;\leq\; \delta\frac{R/2+c}{R}. \tag{3}$$

When the degree of patience of the players is fixed, the property right system must therefore be sufficiently egalitarian to sustain cooperation as a SPE.

Figure 1 provides an illustration of this result. The values of $\phi \in [0,1]$ are read on the x-axis and the relevant values of $\delta \in [0.5,1]$ are read on the y-axis. The two dark lines represent the functions that, for each player, associate to a given share ϕ, the smallest rate of patience δ above which cooperation is better than deviation. $\delta^*(\phi)$ is therefore the highest of these two curves. The horizontal dashed line represents a possible value for the

rate of patience. It crosses the graph of $\delta^*(\phi)$ at the two points of abscissas $1 - \delta\frac{R/2+c}{R} = 0.4$ and $\delta\frac{R/2+c}{R} = 0.6$, that delimit the set of property rights compatible with cooperation under grim strategies.

The situation represented in Figure 1 is therefore one where too much inequality produces a sub-efficient outcome by hindering cooperation. Contrasting with many economics models, that assume *a priori* that incomes reflect contributions to the total output, individuals are assumed here to contribute equally to the common good and are allowed to bargain unequal shares. As a consequence, the poorer individual's contribution is higher than his income. When the wedge between his production and his income becomes too large, he stops cooperating, precipitating the war of all against all. This gives rise to an equality-efficiency complementarity: incomes are higher for both individuals when the property right scheme verifies condition (2).[10]

Importantly, the relationship between equality and efficiency depends on the assumption of "rough natural equality" among agents. When one agent is stronger than the other, in the sense that he is able to obtain a higher income in the war of all against all, equality (i.e. $\phi = 1/2$) does not minimise δ^*, and cannot be said to constitute the social contract most favourable to cooperation (and therefore to efficiency). This can be seen by considering heterogeneous costs in the war of all against all. Assume that agents 1 and 2 have different costs $c_1 \neq c_2$ in the war of all against all. In this case, the property right scheme that minimises δ^* is

$$\text{argmin}_\phi \, \delta^*(\phi) = \frac{R/2 + c_2}{R + c_1 + c_2},$$

which is greater (smaller) than $1/2$ if $c_1 < c_2$ ($c_1 > c_2$). The agent for which the war of all against all is a smaller threat indeed, demands a larger compensation to forego the gain from unilateral deviation. When $c_1 \neq c_2$, the bounds on the distribution of property rights that make cooperation possible are centred around a value different from $1/2$. If equality is not, in this case, the social contract that minimise the likelihood of a war of all against all,[11] the *weak universalisation principle* still holds. An individual

[10]Formally, it is straightforward to check that for any $\delta < 1$, mutual cooperation is a possible SPE only if the state of the world C, C is Pareto improving. That is $\delta_1^{*-1}(\delta) > 1/2 - c/R$, which is equivalent to $\delta_2^{*-1}(\delta) < 1/2 + c/R$. This means that when Inequality 1 is not verified, mutual cooperation is not possible. However in this case, the agent with the highest bargaining power is better-off with a lower share compatible with mutual cooperation. A bargaining procedure with rational self-interested agents would therefore always lead the agent with the highest bargaining power to set his share at the highest level compatible with cooperation. That is, to saturate either side of inequality 3.

[11]Notice that an implicit assumption is that agents know their costs of conflict. The

with a high ability to negotiate property rights and a low cost of conflict is able to extract more resources from the relationship. His share of output nevertheless remains bounded by the other player's willingness to cooperate.

Notice that the assumption of "rough natural equality" is not a moral but a descriptive assumption. In the *Leviathan*, Hobbes states that, in the absence of norms, rules or institutions, individuals are all able to inflict important losses to each other :

> Nature hath made man so equall, in the faculties of body, and mind; as that though there be found one man manifestly stronger of body, or of quicker mind than another; yet when all is reckoned together, the difference between man, and man, is not so considerable, as that one man can thereupon claim to himself any benefit, to which another may not pretend, as well as he. For as to the strength of the body, the weakest has strength enough to kill the strongest, either by secret machination, or by confederacy with others, that are in the same danger with himself. (Hobbes (1985): 183)

The empirical relevance of the "rough natural equality" hypothesis cannot be judged by a casual observation of the current functioning of our societies in which many norms, rules and institutions alter the distribution of abilities to withstand conflict. The theory presented in this paper indeed, does not pretend to capture all the reasons why norms can emerge, but only the reasons that are related to cooperation and that may favour shared progress. Nevertheless norms can also be set (and have often been set in the past) by coercion rather than cooperation; for the sake of personal interest and at the detriment of collective interests. Such norms may provide large advantages to their beneficiaries when it comes to handling conflict. However, such an asymmetry only holds in a world where property rights have already been defined and agreed-upon, and where norms, rules and/or institutions exist to enforce these rights, that is, in a world that cannot be characterised as pre-moral.

The only source of natural inequalities that could influence the cost of conflict in a pre-moral world are physiological characteristics such as strength, intelligence and creativity. If we believe, with Hobbes, that individuals are roughly equal in these respects, then equality is the arrangement most likely

difficulty to assess these costs before choosing to cooperate or defect would, in the absence of a relevant signal, lead both agents to use the expected value of the cost which is, by definition, homogeneous in the population. In the presence of relevant signals, such as gender, size, weight, etc, each agent would condition his cost estimate on the signals.

to support cooperation. If we believe in contrast, that one individual can always dominate the other in a pre-moral context, then equality is not the arrangement most likely to support cooperation (and therefore efficiency). The *weak universalisation principle* is, in this case, more favourable to the agent with the best outside option.[12]

The norm of behaviours induced by the grim trigger SPE has the characteristics of norms of behaviours expected in a contractarian framework. First, it is in the best interest of all agents to adopt them if they are sufficiently *patient*. That is, if they are willing to sacrifice their short-term interest to improve their own long-term well-being, provided that others do the same. Second, the retaliating strategies that players use to prevent defection from the other player captures the *reciprocal* nature of norms of behaviour. Finally, the norm should respect some form of concern for equality (it should at least verify the weak universalisation principle).[13]

Rather than patience, Hobbes relied on the notion of prudence to explain moral principles. In the economics literature, prudence is typically associated with the individual's ability to respond to future risks in an adequate manner while patience reflects the individual's willingness to substitute welfare across time, without any reference to the risky nature of future (and therefore uncertain) outcomes. Even though all future outcomes have elements of uncertainty, it is sometimes possible and useful to disentangle the time preferences from the risk preferences of individuals. The repeated interaction framework, however, makes this disconnection difficult because the discount factor δ may have two interpretations and may be related either to a pure preference for the present in an infinitely repeated game or to a positive probability of continuation in an indefinitely repeated game. This ambivalence of the word patience is well reflected by the title of the seminal Maskin & Fudenberg (1986) paper "The folk theorem in repeated games with discounting or with incomplete information". In practice, since interactions are bound to be repeated a finite number of times, a more natural interpretation of δ is a combination of both a pure preference for the present with a positive termination probability. The notions of patience and prudence hence become tightly intertwined.

[12]Appendix (6.1) offers a complementary explanation. When the stage game is framed in terms of incentives to provide work effort rather than in terms of incentives to preserve peace, the property right system most compatible with cooperation aligns incomes with contributions to total output *in the cooperative state of the world*.

[13]By allowing heterogeneity only in the property rights, and not on the contributions, the current setting only allows to investigate the question of equality, equity being left for further investigation.

3.2 The multiplicity issue

Friedman (1971)'s theorem, along with the "perfect Folk's theorem" of Aumann & Shapley (1976) and Maskin & Fudenberg (1986) among many others, famously show that, as individuals becomes perfectly patient, that is $\delta \to 1$, there exist a continuum of SPEs whose payoffs span the full set of payoffs that are i) (weakly) better than the Nash equilibrium for both players ii) achievable by some strategy. This multiplicity has lead commentators (such as Thoma (2015)) to doubt the relevance of the repeated interaction framework to build a contractarian theory of norms. I challenge this view for three main reasons.

First, the Folk's theorems only hold in the limit, when individuals become perfectly patient. The empirical relevance of this assumption may be challenged since in practice, individuals display limited patience. Whether this limited patience comes from a true preference for the present, or from an understanding of the probabilistic nature of repeated interactions, is an interesting question but it falls outside the scope of this paper. What matters in the context of our analysis is that agents discount future payoffs at a rate that is strictly below one. Many SPE equilibriums may nevertheless co-exist, even for values of δ smaller than one. In particular, the Nash Equilibrium, where no cooperation occurs, is always a SPE. Moving away from the asymptotic analysis that underlies the Folk's theorems therefore mitigates but does not resolves the multiplicity issue

Second, the fact that the theory does not uniquely determine an outcome leaves room for other theories to select which, of the possible equilibirums, actually arises. Evolutionary approaches, in which the wealthier, more cooperative individuals have higher chances of transmitting their memes, provides a strong argument as to why mutual cooperation, if it is a SPE, should emerge as a dominant social norm .

Finally, in a meta-analysis of experimental prisoner's dilemma, Dal Bó & Fréchette (2018) (Fig. 4, p74) show that the proportion of players who cooperate is constant when $\delta \leq \delta^*(\phi)$ and increasing in $\delta - \delta^*(\phi)$ when $\delta \geq \delta^*(\phi)$. This fact does not solve the multiplicity problem but it suggests that $\delta - \delta^*(\phi)$ is a relevant predictor of cooperation.[14]

[14]Using a measure of cooperation $\delta^{RD}(\phi)$ based on Blonski & Spagnolo (2015), Dal Bó & Fréchette (2018) find an even stronger relationship between $\delta - \delta^{RD}(\phi)$ and the rate of cooperation in experiments. In the present model, $\delta^{RD} = \max\left(3/2 - \phi - c/R, 1/2 + \phi - c/R\right)$, which also maximises $\delta - \delta^{RD}(\phi)$ for $\phi = 1/2$.

3.3 A theory of centripetal morality

Using repeated interactions to explain the emergence of norms leads to identify the parameter δ as a crucial factor for the emergence of cooperation. The more patient agents are, and the more often they can be expected to renew their interactions, the more likely is cooperation to be sustained as a SPE. This provides a possible explanation for the emergence of social institutions, such as families, friendships or other networks, that hold individuals liable for a degree of cooperation highly variable according to the frequency of interaction involved. While the very high probability of renewing interactions in the family induces a large incentive to resist short term defections, the lower perceived probability of interacting with foreigners reduces incentives to cooperate.

This view of morality contrasts with Baumard et al. (2013) stating:

> Kin altruism and friendship are two cases of cooperative behaviour that is not necessarily moral. [] In both cases, the parent of friend is typically disposed to favor the offspring or the close friend at the expense of less closely relatives or less close friends, and to favor relatives and friends at the expense of third parties.

The theory presented in the present paper allows the possibility, for a given, potentially mutually beneficial relationship, to harm third parties. Rather than a unique set of universal principles, morality is therefore seen as polycentric, and the jurisdiction of each moral order is defined by its associated cooperative payoffs (R) and likelihood of future interactions δ. A given individual may be part of several moral orders. Relationships characterised by less frequent interactions (lower δ), may sustain cooperation only if they give rise to a more equal split of the surplus or if they entail larger cooperative surpluses. Compared to a family, in which interactions are repeated from one day to the next with a probability very close to one, two individuals in less close interaction are less likely to behave cooperatively (in the sense that the parameter space for which cooperation is a SPE is smaller), unless the joint output of their cooperation is higher (for example because they have complementary skills). The theory hence explains the constitution of concentric circles of moral orders: family, friends, colleagues, neighbours, fellow citizens, humans, living beings, etc, with decreasing intensity of cooperative behaviours as perceived frequency of interaction becomes lower.[15][16]

[15] For the sake of simplicity, the model is written as a two players game. A reformulation of the problem as an n players game is proposed in Section 4.

[16] Despite leaving aside the important question of exploitation, this theory of centripetal morality sheds light on the question of identity. If morality is applied differentially across

This hierarchy of moral orders however, need not be a fatality. Humans have not always been able to cooperate at the scale that we observe today. They have instead discovered, by a process of trial and errors, ways to sustain cooperation in large groups, by identifying the gains from cooperating. The process of technological innovation, permitted by labour specialisation, is an example of such gains. While most individuals rarely interact with researchers, they agree that the benefits from innovation can be considerable. The discovery of new cooperation opportunities can hence bee seen as a major force behind human progress.

4 Cooperation on a large scale

This section extends the previous model to a society of n agents. Our goal is to determine the conditions under which cooperation may arise in a large society. The cooperative surplus $R(n)$, that summarises the amount of wealth created in an harmonious society, is assumed to increase with n, so that larger societies have the potential to create more wealth. Under the rule of law, each individual receives a fraction ϕ_i of the total wealth $R(n)$. When an agent unilaterally defects, he receives an amount $g(n) \leq R(n)$ while other players equally share $R(n) - g(n)$. When several agents simultaneously defect, they share the profit from deviation $g(n)$. Finally, when all players defect, in the war of all against all, each receives $R(n)/n$ and incur a cost c. The grim trigger is now : Defect (D) if one individual defected in the previous rounds and comply (C) otherwise. Cooperation of everybody is possible under these circumstances if and only if no-one is better-off deviating alone, that is

$$\delta \geq \max_i \frac{g(n) - \phi_i R(n)}{g(n) - R(n)/n + c} \equiv \delta(n, \Phi), \tag{4}$$

where $\Phi = (\phi_1, ..., \phi_n)$. As in the two players game, it is the situation of the least well-off individual that determines whether full-scale cooperation is possible or not; and it is straightforward to show that the property right scheme that minimises $\delta(n, \Phi)$, and that consequently maximises the likelihood of cooperation, is such that $\phi_i = 1/n$ for all i.

An important test for the framework presented here is whether it is able to explain cooperation on a very large scale. In order to identify the elements

individuals and groups, it may form a basis for the definition of the notion of identity. The model presented in this paper suggests a plausible causality: repeated interactions create the potential for cooperation which, in turn, may forge the social link required for individuals to identify to the group. The reverse causality may also be true: discovering profitable and repeated joint ventures may be easier in groups where individuals are already united by a feeling of belonging, hence inducing a virtuous circle of cooperation.

of the model that, in addition to equality, allow cooperation to take place, I assume in the remainder of this section that the property right scheme most favourable to cooperation, i.e. $\phi_i = 1/n$ for all i, is in place.

Simply extending the two-players game by setting $R(n) = R$ and $g(n) = R$ and assuming that the property right scheme most favourable to cooperation $\phi_i = 1/n$ is in place, cooperation may occur if and only if

$$\delta \geq \frac{R - R/n}{R - R/n + c} \equiv \delta^*(n). \tag{5}$$

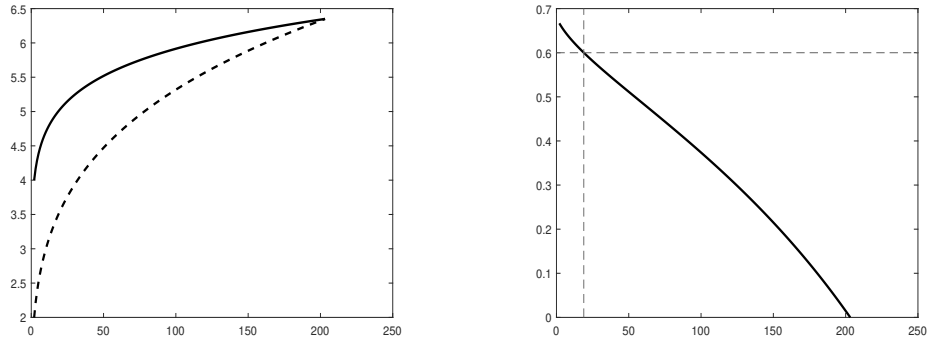
An increase in the number of players raises the relative gain from defection $R - R/n$ (the numerator in Equation 5) by lowering the cooperative share R/n, hence making cooperation harder to achieve, but lowers the payoff in the war of all against all (the denominator), hence favouring cooperation. It is straightforward to see from equation (5) that the former effect always dominates the latter since $\delta^*(n)' > 0$. For a given cooperative surplus R, an increase in the number of players therefore always translates into a less likely cooperative outcome and when $n \to +\infty$, we have $\delta^*(n) \to 1$, even if $R \to +\infty$: no matter how large the cooperative surplus is, cooperation becomes possible in large societies only if individuals are perfectly patient, even though the property scheme most favourable to cooperation is in place.

Individuals may therefore cooperate on large scales only if the gain from unilateral deviation $g(n) - R(n)/n$ decreases with n, in which case we have $\delta^*(n)' \leq 0$. Equation (4) shows that, if the deviation payoff $g(n)$ grows at a lower pace than the cooperative payoff $R(n)/n$, then $\delta^*(n)$ may indeed decrease with n. Figure 2 illustrates an example of such a society.[17] The left-hand side panel represents the gain from deviation $g(n)$ (full line) and the gain from cooperation $R(n)/n$ (dotted line). The gain from cooperation $R(n)/n$ is assumed increasing in n, to capture potential gains from specialisation or other economies of scales only attainable in larger groups. Other things held constant, these additional gains make cooperation more difficult to sustain since a defection allows individuals to capture a larger output.[18] If however, the group is able to control the gain from deviation, such that $g(n)$ increases with n at a lower pace than the gain from cooperation, as is represented on the left-hand side panel of Figure 2, then cooperation may become possible in large groups. The right-hand side panel shows that, despite the increase of the gain from deviation with the number of individuals, the patience threshold, above which cooperation is possible, decreases with n. In this calibrated example, a level of patience $\delta = 0.6$ results in cooperation only being possible in a group with more than 19 people. Of course,

[17] The calibration is reported in Appendix 6.2.

[18] In mathematical terms, if $g(n) = R(n)$, then $\delta^*(n)' > 0$.

since the gain from unilateral deviation grows at a lower pace than the gain from cooperation, the game is not a prisoner's dilemma when n is very large ($n \geq 204$ in the example represented in Figure 2) and cooperation becomes a Nash Equilibrium of the stage game, and therefore a SPE for any value of patience δ.



(a) Functions $g(n)$ and $R(n)/n$ (b) Function $\delta^*(n)$

Figure 2: The left-hand side panel represents the function $g(n)$ (full line) and $R(n)/n$ (dotted line). $R(n)/n$ is lower than $g(n)$ but increases faster. The function $\delta^*(n)$, represented on the right-hand side panel, is therefore increasing in n. For a rate of patience $\delta = 0.6$, cooperation only becomes a SPE when $n \geq 19$. The calibration is given in Appendix 6.2.

While too much inequality within the property right scheme remains a threat to cooperation, this simple example shows that large societies must also control the gain from a unilateral deviation, which is a kind of inequality between those who comply and those who deviate. Even though the mechanism through which this is achieved remains outside the scope of this paper, it shows that the theory presented is able to account for the institution of norms of behaviours in very large groups, where individuals interact at a low frequency, but where the gains from cooperation can become very large.

5 Non-grim strategies

The grim trigger involves an extreme form of retaliation and since no harsher punishment can be adopted in the prisoner's dilemma than an infinite sequence of Ds, the grim trigger is the one that gives the most chance to cooperation actually arising. Cooperation in the indefinitely repeated prisoner's dilemma is therefore possible only if cooperation through grim strategies is

possible. Condition (3) can hence bee seen as a minimal condition for cooperation. However, many other strategies are also SPE for a given value of patience δ. Players can, for example, adopt strategies that punish a deviating player a limited number of times or play alternating sequences of C and D.[19]

Nevertheless, as long as the strategies have symmetrical punishments, the same result can be obtained : the more equal the distribution of payoffs, the more likely they are to be sustained as SPE. The same reasoning as before applies here: no matter what payoffs a given pair of strategies provides, the player with the lowest gain from cooperation needs to value the future more than the other player to be willing to cooperate. Increasing this player's share of the cooperative surplus hence results in a lower threshold δ^* above which cooperation is a SPE. This reasoning however, does not apply to strategies that involve asymmetric punishments.

Formally, the fact that grim strategies are the most susceptible to give rise to cooperation can be expressed as follows. Let $\delta^*_{GT}(\phi)$ be the patience threshold above which cooperation is a SPE when the two players adopt a grim trigger, and $\delta^*_S(\phi)$ be the patience threshold when at least one player plays another strategy. Then for any such pair S of strategies, we have

$$\min_{\phi} \delta^*_{GT}(\phi) = \delta^*_{GT}(1/2) \leq \min_{\phi} \delta^*_S(\phi). \tag{6}$$

The proof of this statement is rather straightforward but it requires considering two cases. To be part of a SPE, a cooperative strategy must entail a punishment, for there is otherwise no incentive for the other player to comply (play C). Nevertheless, the two strategies may be asymmetrical, with one player being more severe than the other in case of deviation. Let us first consider the symmetrical case. Since individuals are assumed identical except

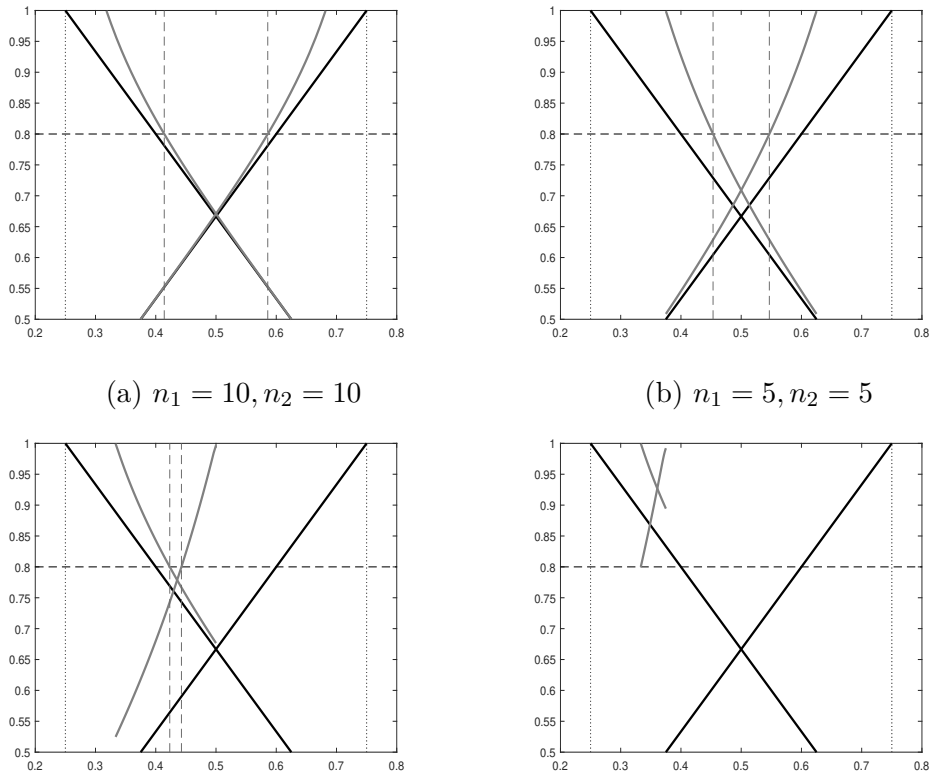
[19]For example, players may take turn playing D while the other plays C. Such cooperating strategies allow to modulate the frequency of each agent playing D and hence receiving the high payoff. Compared to an infinite sequence of (C, C), that gives each players his share ϕ and $1 - \phi$ of the cooperative output R, this allows to modify the distribution of payoffs. In the limit, when $\delta \to 1$, such strategies allow to span to full set of feasible and individually rational allocations and become all sustainable as SPE. Players can also choose to play behavioural strategies, in which they define a probability of playing C and D, conditionally on the history of the game. Allowing such strategies requires a public randomisation device, permitting each player to identify deviations from his opponent/partner and to react accordingly. The existence of such a device in practice is disputable but from a theoretical perspective, it allows the players to reach any feasible and individually rational payoff, even when $\delta < 1$ is fixed. Notice that, allowing agents to use such a randomisation device, the parameter ϕ can be interpreted as a mixing probability of the pair of strategies where the row player plays D and the column player plays C when a given event of probability ϕ occurs, and where the row players plays C while the column players plays D when the complementary event occurs.

in terms of property rights endowment (ϕ), it is the situation of the least well-off that determines the cut-off δ^* equalising the discounted gains from defection to the discounted gains from cooperation. When $\phi \neq 1/2$, increasing the share of the least well-off individual therefore results in a decrease in δ^*. Hence $\text{argmin}_\phi \delta^*_{GT}(\phi) = 1/2$. In addition, $\delta^*(\phi)$ is also a function of the punishment length. If two symmetrical strategies entail only a finite number of punishment periods (playing D after observing the other player having played D), then increasing the number of punishment periods reduces $\delta^*(\phi)$ for any $\phi \in (0, 1)$ and in particular $\delta^*(1/2)$. Since the pair of grim strategies entails an infinite number of punishment periods, no other pair of strategies with symmetrical punishments has a lower $\delta^*(1/2)$.

Asymmetrical punishments lead us to consider cases where one agent punishes the other for longer periods of time after observing a deviation. Such a pair of strategies can be a SPE as long as neither player has an incentive to deviate. Under these circumstances, setting $\phi = 1/2$ does not minimise δ^*. Indeed, for a given ϕ, the agent who faces the lightest punishment has more incentive to deviate. Setting $\phi = 1/2$ therefore necessarily results in this agent having a lower patience threshold than the other agent, who faces the harshest punishment. Since cooperation is SPE only if both agents cooperate, increasing the share of the player who faces the light punishment reduces the threshold of the game δ^*. Nevertheless, increasing the length of the punishment period for the agent with the lightest threshold always results in lowering the threshold δ^*. And as both punishment periods become infinitely long, the pair of strategies considered converges to a pair of grim triggers, for which $\text{argmin}_\phi \delta^*_{GT} = 1/2$.

Cooperation can be achieved more easily in a situation of imperfect equality ($\phi \neq 1/2$) when the punishment is asymmetric, however, inequality (6) shows that the pair of grim triggers with egalitarian property rights is the social contract most likely to provide the efficient cooperative output.

For a given level of patience δ, this does not, of course, rules out asymmetric strategies. In these cases, where imperfect equality is more likely to give rise to cooperation, bounds on the maximum inequality (weak universalization principle bounds) still constrain the set of efficient cooperative outcomes and it is always in the best interest of both players to respect these bounds. A player with a very strong bargaining ability would therefore willingly limit his power in order to preserve the efficient cooperative outcome.



(a) $n_1 = 10, n_2 = 10$ (b) $n_1 = 5, n_2 = 5$

(c) $n_1 = 2, n_2 = 8$ (d) $n_1 = 1, n_2 = 8$

Figure 3: The values of $\phi \in [0, 1]$ are read on the x-axis and the relevant values of δ are read on the y-axis. The highest of the two grey lines represents the function $\delta^*(\phi)$. The horizontal dashed line represents a possible value (0.8) of the rate of patience, crossing the graph of δ^* at the two points that define the largest inequality compatible with (C, C) being a SPE. The computations were realised with $R = 4$, $c = 1$ and $\delta = 0.8$. Each graph is associated with a pair of (possibly asymmetric) trigger strategies, where player 1 imposes a n_1 periods punishment and player 2, a n_2 periods punishment in case of deviation.

Figure 3 illustrates the effect of a change in the punishment duration on the maximal inequality that allows cooperation to take place. As previously, the black lines represent the smallest discount factors above which cooperation is possible if both agents use a grim trigger. The grey lines represent the same critical thresholds when agents use trigger strategies with finite punishment lengths (cooperate as long as the other player cooperate and defect for n_i periods when player $j \neq i$ defects). Sub-figure 3a that illustrate the symmetric strategy where both agents retaliate for $n_1 = n_2 = 10$ periods in

case of defection by the opponent. Compared to the pair of grim triggers, the strategy with lower punishment yields higher levels of critical patience and therefore tighter bounds on the property right scheme. Choosing shorter punishment duration increases even more the tightness of the bounds, as illustrated in sub-figure 3b for the case $n_1 = n_2 = 5$. Sub-figure 3c and 3d illustrate the case of asymmetric punishments. On sub-figure 3c, player 1 punishes a shorter period $n_1 = 2$ than player 2, who reverts to D for $n_2 = 8$ periods after observing a deviation from player 1. In this setting, a defection entails a higher cost for player 1 than for player 2. Hence, for a given level of patience, the lowest share of output that player 1 is willing to accept is lower than the lower share of output that player 2 is willing to accept. In other words, player 1 has more to loose from deviating than player 2, which results in a translation of the weak universalization principle bounds. An asymmetry in the ability to punish therefore potentially translates into inequality in the property right scheme. Sub-figure 3d, however, shows that when punishments becomes too asymmetric, cooperation may become incompatible with the actual level of patience δ of both agents. In such cases, cooperation never takes place and both agents remain in the war of all against all state.

6 Conclusion

This paper uses indefinitely repeated game theory to generate endogenous moral norms. Under Hobbes' assumption of "rough natural equality", reciprocity and equality are shown to favour cooperation and efficiency. Under more general assumptions, a condition that reproduces Moehler (2018)'s *weak universalisation principle* is obtained in a two players prisoner's dilemma. The paper hence develops a theory of self-interested morality, in which *homo patient* agrees to constrain his own behaviour, hence sacrificing his short-term interest in order to preserve a mutually beneficial cooperation in the long run. A theory of centripetal norms is sketched, in which moral norms are applied differently across groups, depending on the likelihood with which interactions can be expected to occur in the future. The role of norms, as a mean to achieve cooperation and efficiency, is then illustrated in large groups, in which the "free-rider problem" may be circumvented if the gain from unilateral deviation increases with the number of players at a lower pace than the gain from cooperation. This remark opens the door to the fascinating question of how institutions may complement moral norms to control the "free-rider problem". If this question deserves to be treated in more depth, the theory presented here suggests that formal institutions may be needed in large groups to curve the free-riding issue, by limiting the gains

from violating norms. This delegation of power however, is coherent with the contractarian framework only if it results from the direct expression of individual's will to constrain their own behaviours provided that everyone's behaviours is similarly constrained. Finally, the paper studies the robustness of the *weak universalisation principle* to situations where individuals play non-grim strategies. This test is important because such strategies can support cooperation in the indefinitely repeated games and have *a priori* no fewer reasons to arise than the grim trigger. However, I have argued that grim triggers are central because i) they are the ones where cooperation is most likely to arise due to their harsh punishment that deter unilateral deviation ii) any *weak universalisation principle* bound on property rights of a non-grim pair of strategies is comprised within the *weak universalisation principle* bounds of the grim trigger pair.

The theory of efficient norms presented here makes a case for morality, and in particular for equality, as a mean of common progress and advancement. It does not, however, pretend to be a comprehensive positive theory of norms and is therefore fully compatible with the existence of norms imposed by coercion in the interest of the few and at the detriment of the many. Investigating the condition of existence and perpetuation of such norms and their interactions with the type of efficient norms presented in this paper is a fascinating topic left for further research.

6.1 Work-Shirk interpretation of the game

The game was presented so far in terms of incentives to respect the established order. However, a different presentation of the game allows to frame all results obtained above in terms of incentives to provide effort. Assume that the players can produce an output R. If they both work, they incur a personal cost of effort e, but if only one agent exerts effort, the cost to him is $E > e$. There is therefore an incentive to free-ride on the other player's effort. If neither of the agents works, the output is 0. For simplicity, I assume that it is never worth doing the effort alone, so $R < E$. As previously, the respective shares ϕ and $1 - \phi$ of the total output (now either R or r) earned by the agents are decided through an exogenous bargaining procedure before the game starts. Table 2 summarises the game. The situation (S, S), where

2^*		W	S
	W	$\phi R - e, (1-\phi)R - e$	$\phi R - E, (1-\phi)R$
	S	$\phi R, (1-\phi)R - E$	$0, 0$

Table 2: Stage work-shirk game

both agents shirk, is the unique Nash Equilibrium in the stage game.

Similarly to the game presented in Section 2, the Nash Equilibrium of this game is Pareto dominated by the state where both individuals work if the property right scheme is not to unequal. In particular, the condition under which the game is a prisoner's dilemma is

$$\frac{e}{R} < \phi < 1 - \frac{e}{R}, \tag{7}$$

which is possible for some value of ϕ if and only if $2e < R$. The assumption $2e < R$ means that the average cost of producing R is lower when both agents work. This may be due to increasing individual marginal cost of effort, a common assumption in the economics literature, or to synergies between workers that allow them to be more efficient when they work together.

Under condition 7, the game is therefore a Prisoner's dilemma and cooperation is a SPE of the repeated game if and only if

$$\frac{e}{\delta R} \leq \phi \leq 1 - \frac{e}{\delta R}.$$[20]

Too much inequality in the property right scheme therefore harms cooperation and efficiency.

[20]Since $\delta \leq 1$, cooperation is therefore possible only if condition 7 holds.

6.2 Cooperating in large games - Calibration

Let $R(n)$ be a function with constant elasticity such that $R(n) = k_1 n^\sigma R$, with $\sigma > 1$. In order for the n players game to be an extension of the two players game k_1 is calibrated such that $R(2) = R$, which gives $k_1 = 2^{-\sigma}$. I also assume

$$g(n) = \frac{R(n)}{k_2 n^\psi},$$

where $\psi \geq 1$ captures the ability of society to deter free-riding and k_1 is a constant. We therefore have $g(n) = k n^{\sigma-\psi} R$, where $k = k_1/k_2$. k is calibrated so that $g(2) = R$, as in the two players game, which implies $k = 2^{\psi-\sigma}$. $g(n)$ is increasing if and only if $\sigma \geq \psi$ and $\delta^*(n)$ is decreasing in n if and only if

$$n \geq 2^{\frac{\psi}{\psi-1}} \Big(\frac{\sigma-\psi}{\sigma-1}\Big)^{\frac{1}{\psi-1}}.$$

When $\sigma < 2\psi - 1$, $\delta^*(n)' < 0$ for any $n \geq 2$. It is therefore possible to have $\delta^*(n)' < 0$ (cooperation becomes more likely as n grows) despite $g(n)' > 0$ (the gain from deviation increases) when

$$\psi < \sigma < 2\psi - 1,$$

which corresponds to the situation represented on Figure 2. The precise example depicted uses $\sigma = 1.3$, $\psi = 1.15$ and $R = 4$.

7 Bibliography

References

Aghion, P., Caroli, E., & Garcia-Penalosa, C. (1999). Inequality and economic growth: the perspective of the new growth theories. *Journal of Economic literature*, 37(4), 1615–1660.

Aumann, R. & Shapley, L. (1976). Long term competition. *A Game Theoretic Analysis*.

Axelrod, R. (1981). The emergence of cooperation among egoists. *American political science review*, 75(2), 306–318.

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(1), 59–78.

Berg, A., Ostry, J. D., Tsangarides, C. G., & Yakhshilikov, Y. (2018). Redistribution, inequality, and growth: new evidence. *Journal of Economic Growth*, 23(3), 259–305.

Blonski, M. & Spagnolo, G. (2015). Prisoners' other dilemma. *International Journal of Game Theory*, 44(1), 61–81.

Dal Bó, P. & Fréchette, G. R. (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, 56(1), 60–114.

Friedman, J. W. (1971). A non-cooperative equilibrium for supergames. *The Review of Economic Studies*, 38(1), 1–12.

Fudenberg, D. & Tirole, J. (1991). Game theory, 1991. *Cambridge, Massachusetts*, 393(12), 80.

Gaus, G. (2019). Moral conflict and prudential agreement: Michael moehler's minimal morality. *Analysis*, 79(1), 106–115.

Gauthier, D. (1986). *Morals by agreement*. Oxford University Press on Demand.

Hobbes, T. (1985). Leviathan. macpherson cb, ed.

Maskin, E. & Fudenberg, D. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 53(3).

Messina, J. & Wiens, D. (2020). Morals from rationality alone? some doubts. *Politics, Philosophy & Economics*, (pp. 1470594X20906616).

Mirrlees, J. A. (1971). An exploration in the theory of optimum income taxation. *The review of economic studies*, 38(2), 175–208.

Moehler, M. (2018). *Minimal morality: A multilevel social contract theory*. Oxford University Press.

Moehler, M. (2020). Contractarianism.

Piketty, T. & Saez, E. (2013). Optimal labor income taxation. In *Handbook of public economics*, volume 5 (pp. 391–474). Elsevier.

Thoma, J. (2015). Bargaining and the impartiality of the social contract. *Philosophical Studies*, 172(12), 3335–3355.