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*

March 3, 2021

Abstract: Building on the contractarian tradition, I propose a two-step bargaining theory of justice. Individuals first bargain about the distribution of property rights over total wealth. They subsequently play an indefinitely repeated prisoner's dilemma in which they either comply with the property right system previously decided, or they fight to increase their share. Mutual defection leads to a "war of all against all" while mutual cooperation generates a peaceful and efficient outcome. I show in this framework that mutual cooperation can be an equilibrium only if inequality is sufficiently low. In such a case, an individual with a high bargaining power restricts his claim over total wealth in order to preserve cooperation. This strategic behaviour is observationally equivalent to a purely altruistic behaviour. The model hence offers a possible rationalisation of individual's demands for equality and justice.

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.

An important goal for the contractarian theory, and more broadly for bargaining theories of justice, is to explain how the moral status of the individual and his aspirations for equality, rightfully acknowledged in important theories of justice, such as Rawls (1971), Sen (1980), Dworkin (1981), Roemer (1998) and Anderson (1999), come to arise.[1]

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 power. Once production has taken place, each agent

[1]The necessity of this aspiration for equality and morality in the theory of Rawls (1971) is demonstrated in Moreno-Ternero & Roemer (2008).

may either respect the existing property rights or dispute them in view of obtaining a larger output share. Dispute by both players leads to a value destruction, akin to Hobbes' war of all against all, where individuals waist large amounts of resources because of conflict.[2] In contrast, peaceful cooperation prevails if both agents respect the property rights previously decided. Cooperation may arise in the indefinitely repeated version of this prisoner's dilemma only if individuals are sufficiently patient and if their probability of renewed interaction is sufficiently high. In this framework, the Folk's theorems famously show that individual rationality is compatible with *any* distribution of income, hence drastically restricting the predicting power of repeated game theory. This limit case however, should not obfuscate the fact that real people display limited patience and can never be expected to renew their interactions from one period to the other with absolute certainty. If game theory does not provide a unique prediction as to what may happen in this case, it nevertheless allows to reject certain outcomes. In particular, I show that agents are able to cooperate in the indefinitely repeated prisoner's dilemma only if the distribution of the final income is not too unequal and if agents adopt reciprocal behaviours. Since both agents are better-off under cooperation, 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 (apparently) altruistic behaviour is obtained by disentangling individual's short-term and long-term interests. In the short run, individuals should restrain their bargaining power in order to preserve their long-run interests. More specifically, the maximal inequality condition formalises Moehler (2018)'s *weak universalisation principle*: at all time, each individual must make sure that others receive a sufficient fraction of the cooperative output to satisfy their basic needs. Once this condition is satisfied, each individual is free to make use of his bargaining power. The model presented 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

[2]Notice that Hobbes' expression of a war of all against all need not be understood literally. What matters in our context is that both individuals are worse-off when they do not cooperate and reject the property right scheme.

worse outcome than compliance with the principle.

Equality is therefore shown to be a cooperation-enabling device that society must use in order to reach efficient outcomes. In line with the democratic egalitarianism developed in Anderson (1999), this result owes much to the recognition that a modern economy should be regarded as "a system of cooperative, joint production", where every product incorporates a vast number of individual inputs, rather than "a system of self-sufficient Robinson-Crusoes, producing everything all by themselves" (Anderson, 1999, p. 321). The model also offers a response to the vulnerability criticism, according to which contractarian theories fail to grant moral standing to severely disabled and dependent individuals. In a society where labour is highly specialised, the inter-dependence between individual's contributions implies that abilities cannot be attributed to individuals alone. Instead, abilities are relative to a given societal organisation that can be modified to value each individual's set of intrinsic abilities. When such a re-organisation of the social contract is feasible, it is in the best interest of all individuals to do so. Relatedly, the model allows for a simple articulation of the positive and normative aspects of justice: while a positive perspective accommodates the lower moral status of individuals unable to participate to cooperation *within the current social contract*, a normative perspective may recommend the adoption of new forms of societal organisation. From both perspectives, the proposed framework also allows to derive implicit weights, that the norm of justice grants to each individuals, without resorting to inter-personal comparisons of welfare. Indeed, since individuals are assumed self-interested, they only compare their own current and future welfare.[3]

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 present paper shows that moral constraints can be derived in a contractarian framework on the sole basis of non-moral assumptions. Along with the representation of society as a "system of joint production", Hobbes' "rough natural equality" assumption is shown to play an important role.

[3]This does not mean that individuals cannot have other-regarding preferences. The simplest interpretation of the model involves non-tuistic preferences but all results and interpretations may also support a tuistic preference assumption.

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 strategies consistently out-preform others, this literature explains the emergence of norms as dominant social practices. The present paper remains up-stream of this literature and shows that too much inequality prevents the emergence of cooperation among rational agents. Nevertheless, the fact that cooperation can happen does not mean that it has to happen.[4] An evolutionary argument is therefore put forward to explain why 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.[5] 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). While it shares with Binmore et al. (1994) a common objective to ground the concept of justice within a positive framework, two major differences should be highlighted. First, the veil of ignorance is never used as an input of the model. It is nevertheless found to be an appropriate heuristic since in the cooperative equilibrium, individuals restrict their claims on the total output, hence behaving as if they were putting themselves in other people's shoes. Second, the analysis allows for a characterisation of both the positive and normative dimensions of justice.

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

[4]The equilibrium concept used in this paper is Subgame Perfection, which notoriously yields multiple outcomes. In particular, the Nash Equilibrium where agents fail to cooperate is 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.

[5]The model of competition between groups that underlies this argument is not explicitly modelled in this paper but a simple version can be summarised as follows. Assume that two societies, identical in all respect, use resources to compete against each other. Since, in our setting, the richest society is the one that enforces internal cooperation, the norm of cooperation, supported by the equality condition, has a higher chance of survival than the norm of defection.

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.[6] 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 bring in the laboratory. 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,[7] or 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. The model presented here is not incompatible with the theoretical underpinnings of these models but it offers a counterbalancing mechanism, through which equality may be a source of efficiency.

This mechanism 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.[8]

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. The main assumptions of the model along with the vulnerability and multiplicity

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

[7]This is the case of the optimal taxation literature initiated by Mirrlees (1971). For a more recent review of the literature see Piketty & Saez (2013).

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

issues are discussed this section. 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 restriction 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, p. 12) in the war of all against all state of nature.[9][10]

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

[9]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.

[10]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.

let the other player unilaterally dispute the property right scheme, hence reaping the full cooperative output.

	C	D
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 property right scheme 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

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

both agents are better-off if they comply with the property right scheme.[11] Contrasting with Moehler (2018)'s and Messina & Wiens (2020)'s representation of the Hobbesian framework, the game summarised in Table 1 is, in this case, a prisoner's dilemma.[12]

Individual rationality leads both agents to defect even though it would be profitable for them to cooperate (given that the other player cooperates),

[11] 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 mutual cooperation, 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 16.

[12] The choice of representing interactions as a prisoner's dilemma is not uncontroversial. Vanderschraaf (2006) follows Locke in assuming that the fact that "a rational individual prefers to Anticipate against another only when she expects the other to Anticipate" (p 250-251) leads to an Assurance Game structure rather than a Prisoner's Dilemma. This Lockean assumption may indeed represent some situations that are less conflictual than the one presented in this paper, for example because individuals care for each other or because it is not possible to get away with unilateral defection. Such a set-up however, seems less likely in a pre-moral society with no institutions. Even our current societies remain plagued with such opportunities when institutions and moral norms fail to provide the adequate incentives. The problem of climate change and our inability to curb carbon emission is an important illustration of such institutional failures.

because none of them can be sure that the other will indeed cooperate. Traditional analysis of the Hobbesian framework (for example Moehler, 2009) labels this issue the assurance problem. It is conceptually kept separated from the compliance problem that arises because of individual's lack of commitment power. Indeed, once the assurance problem is solved and a social contract is agreed-upon, individuals also have to be bound to their agreement.

To solve both problems, Hobbes concluded in his second and third laws of nature, that individuals should give-up their natural rights 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, p. 20). In addition, this authority should be allowed to "threaten agents with severe sanctions for defective behaviours" (Moehler, 2020, p. 14). 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 "alienation contract" (Hampton, 1986) solves both the assurance and the compliance problems but 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 powerful 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, 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.[13]

3 Indefinitely repeated games

The two players now 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

[13]Even though norms and formal institutions may, in practice, have the same purpose of maintaining cooperation and if institutions have to be legitimated by individuals in the same way that norms do.

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 in equilibrium.

3.1 Grim strategies

The pair of grim strategies is a Subgame Perfect Equilibrium (SPE) if and only if none of the players have a profitable deviation when they play C,[14] which may be written mathematically for the row player as

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

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

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

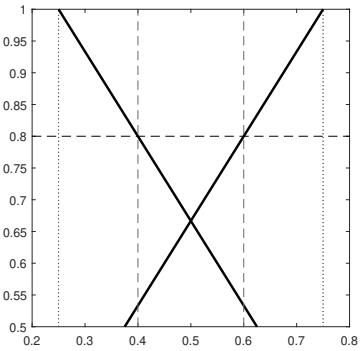
gives the condition under which deviation is not a profitable strategy for the column player.

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

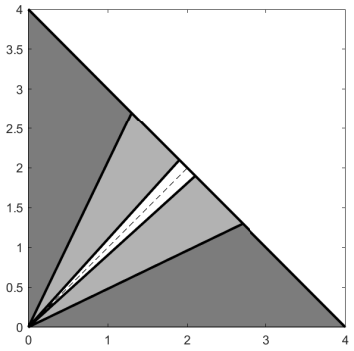
$$\delta \geq \max\Big(\frac{R(1-\phi)}{R/2+c}, \frac{R\phi}{R/2+c}\Big) \equiv \delta^*(\phi), \tag{2}$$

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 range of patience values $[\delta^*(\phi), 1]$, compatible with cooperation, 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. Or, the long-term gain from cooperation increases with the cooperative surplus, while the short-term gain from deviation decreases. For a given player, an increase in his cooperative share therefore provides more incentives to cooperate. 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.



(a) Discount factor $\delta^*(\phi)$



(b) Cooperation set

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. Condition (3) is a formalisation of Moehler (2018)'s weak universalisation principle.[15]

Figures 1a and 1b illustrate these results. Figure 1a depicts the set of patience values $\delta \in [0.5, 1]$ compatible with cooperation as a function of the values of $\phi \in [0, 1]$. 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

[15]Several differences with Moehler (2018)'s theory must nevertheless be acknowledged. First, Moehler (2018)'s is a two-stage theory of morality where individuals begin by using their shared moral norms and values to settle the most superficial conflicts. They resort to the kind of instrumentally rational morality described in this paper only in case of deep moral disagreement, when the conflict bears on the moral values themselves. Second, Moehler (2018)'s weak universalisation principle is defined directly in terms of basic needs. In contrast, the bounds obtained here do not make explicit reference to individual's needs. If we consider that individuals have biological needs that they must meet in order to survive, then such needs trivially defines an additional equality constraint for the game. Alternatively, we may consider that inequality 2 contributes to define individual's basic needs. In this case, what individuals need is to have a prospect of interaction that is (significantly) better than the prospect of a lifetime of conflict with their peers.

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. For the sake of this illustration, the computations were realised with $R = 4$, $c = 1$ and $\delta = 0.8$. Figure 1b presents the same analysis in the more usual space of individual's final incomes. The feasible set is delimited by the function $f(x) = R - x$ and the Folk's theorems predict that any outcome within this feasible set may obtain as $\delta \to 1$. In contrast, when $\delta = 0.9$, the darker grey area can never be reached and when $\delta = 0.7$, the lighter grey area must also be excluded from the set of potential final payouts. Indeed, the smaller the level of patience, the more difficult it is to sustain cooperation and therefore, the tighter the inequality condition (3). The situation represented in Figures 1a and 1b 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).[16]

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

[16]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 never lead the agent with the highest bargaining power to set his share above the highest level compatible with cooperation given by inequality (3).

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 $c_1 < c_2$ ($c_1 > c_2$). Indeed, the agent for which the war of all against all is a smaller threat 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,[17] the *weak universalisation principle* still holds. An individual 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.

It is worth noticing that, even in the case of asymmetric costs of conflict, δ^* increases with R and tends to one half when $R \to +\infty$. Similarly, the bounds on ϕ such that cooperation is possible shrink as R grows. This means that when the cooperative output becomes more important, equality becomes a more stringent constraint, no matter how unequal the situation in case of conflict.

Importantly, 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, p. 183)

[17] An implicit assumption is that agents know their costs of conflict. The 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.

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 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.[18]

The norm of behaviour induced by the grim trigger SPE has the characteristics 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. Finally, the norm reflects a concern for equality since it should at least verify the weak universalisation principle.

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 ele-

[18]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*.

ments 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.[19]

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 $\delta \to 1$,

[19] Another conceptual ambivalence can be sorted out by acknowledging Moehler (2009)'s suggestion that Hobbes' third law of nature is best captured in the language of game theory by the use of a grim strategies. Moehler (2009) argues that Hobbes' assurance dilemma is best represented by a one-shot assurance game, in which individuals have to coordinate on the best of two possible Nash equilibria. Once they have solved this problem by instituting a sovereign, the latter must threaten potential violators of the social contract with a severe threat (in this sense, the sovereign follows a grim strategy) to deter short-sighted individuals from deviating in the assurance game. Since the present model dispenses with Hobbes' sovereign, individuals solve both the assurance and the compliance problem by playing grim strategies with each other even though, according to Moehler (2009) (p323),

> under the fierce conditions descried by Hobbes, an individual cannot expect to interact repeatedly with her fellows in the state of nature[.]

The present model in its simplest form however, assumes that individuals may play indefinitely even when they fail to cooperate and end-up in the war of all against all state of the world. That is, it does not considers literally the concept of the war of all against all. It is however straightforward to adapt the framework in order to take more seriously the threat of termination in the case of conflict. Indeed, we may assume that when both agents defect, the continuation probability of the game is $\underline{\delta} < \delta$. In this case, mutual cooperation is possible if

$$\frac{1-\delta}{1-\underline{\delta}}[1-\underline{\delta}(1/2+c/R] \leq \phi \leq 1-\frac{1-\delta}{1-\underline{\delta}}[1-\underline{\delta}(1/2+c/R],$$

which is a weaker constraint than (3) because the threat of termination provides additional incentives to cooperation. All the results of the paper go through with this modification.

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 sometimes leads to the conclusion that the repeated interaction framework is unsuited 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 resolve 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, p. 74) 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.[20]

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

[20]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 $\phi = 1/2$.

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 or 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 allows for 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.[21][22]

This hierarchy of moral orders however, need not be fixed in time. Humans have not always been able to cooperate at the scale that we observe today. They have instead discovered, by trial and errors, ways to sustain

[21] 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.

[22] 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 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.

cooperation in large groups. The process of technological innovation, credited for the large increases in life standards across the globe, could not have taken place without labour specialisation, which is nothing else but a vast system of cooperation. The complexity of this system makes it hard to disentangle individual contributions to the cooperative output. In our globalised economies, a researcher who discovers a new vaccine against a lethal and contagious disease can hardly be credited alone for his discovery. Indeed, the social contract that allows him to spend his time doing research is only made possible by the thousands of people whose activities consist in providing him with food, cloth, safety, education, health care, etc. This deep interconnection between individual contributions is easy to forget because each of us only has the ability to consciously acknowledge a limited number of interactions. The scale at which our societies practice labour specialisation is therefore only possible at the cost of an anonimisation of these interactions, which makes them difficult to see. Our researcher cannot and does not have to acknowledge the person who drives the bus that takes him every morning to work. Nor does he have to acknowledge the people who tailor his cloth in factories located on the other side of the globe. Instead, everyone is given a compensation for his work, that we may only hope to reflect principles of justice. However, there is no way of assessing the fairness of such a distribution of compensations in a pre-moral context.

3.4 Asymmetric productivity

I have assumed so far that both individuals contribute equally to the total cooperative output R. This assumption is motivated by the difficulty to identify individuals' contributions to total production in our highly specialized societies. Unless one believes that actual markets are perfectly competitive, in which case the contribution of each is reflected by his market income, all one can hope for is that market incomes offer at least some proxy for contribution. Indubitably, this information would be of great help to design a fair system, promoting effort and cooperation. Nothing however, allows such a conclusion at a general level, for we know many markets are noncompetitive and some are nonexistent. Second-best theory also implies that lack of competitiveness in even a single markets may trickle down to other markets and result in non-competitive outcomes in all markets. As a consequence, there is no way of saying whether actual markets incomes indeed proxy for contributions.

There are however, situations in which individuals may be able to rank and value their respective contributions. In a two players game, they may even agree that one player is more essential to cooperation than the other.

This may be represented by an asymmetric gain. Assume that agent 1 can produce an output of value R_1 for himself without the help of agent 2. Similarly, agent 2 can produce an output of value R_2 without the help of agent 1. When $R_1 > R_2$, we can say that agent 1 contributes more than agent 2 because his cooperation brings more to agent 2 than agent 2's cooperation brings him. This case may represent a situation where the joint venture is a physical task, for which agent 1 is more efficient than agent 2. The stage game is represented in Table 2.

	C	D
C	$\phi R - R_1, (1-\phi)R - R_2$	$-R_1, R - R_2$
D	$R - R_1, -R_2$	$R/2 - R_1 - c, R/2 - R_2 - c$

Table 2: Stage game, with asymmetric payoffs

The inequality threshold that defines the possibility of cooperation is unmoved in the asymmetric contributions case and still given by inequalities (3). This is because when players adopt grim-trigger strategies, they only care about the difference between their personal gains from cooperation and from deviation, which is unaffected by their outside options. How the other person benefits from cooperation plays no role in the decision to cooperate.

The intuition that people who contribute more should receive more can be recovered from the participation constraints of the agents. Assume indeed that individuals have the possibility not to play the game, in which case they receive their outside options R_1 and R_2. Assuming that the condition for cooperation (3) holds and that individuals indeed succeed to cooperate, participation of both individuals requires

$$\frac{R_1}{R} \leq \phi \leq 1 - \frac{R_2}{R}. \tag{4}$$

Constraint (4) is symmetric only if both agents have the same outside options. When one individual has a higher outside option, he may require a higher compensation to participate to the game. Depending on the values of R, R_1 and R_2, condition (4) may be binding or not. The higher the outside options of the players, the more likely binding is condition (4). In contrast, the higher the gains from cooperation, the looser (4) becomes and when $R \to +\infty$, condition (4) is not a constraint anymore.

The relevance of the participation constraint depends on the assumptions one makes on individual's ability to leave society. If one believes that agents can earn on their own a significant amount of what they earn in society, then

the participation constraint is relevant and may impose a strong inequality in the distribution of incomes, to the extent that individuals are able to identify relative contributions to the cooperative output. In contrast, if agents have no ability to live on their own, or to join other communities, then the participation constraint should not affect the distribution of goods.

3.5 Vulnerability : from positive to normative

Contractarianism is sometimes criticised for providing moral standing only to those who are able to contribute to the joint production process, leaving aside the most vulnerable such as children and disabled people. If a thorough discussion of the vulnerability issue is outside the scope of this paper, the present framework allows us to present two arguments against this criticism. First, if it is understood that anyone can, with some positive probability, be in the situation where he cannot contribute to the joint output, then it may be rational for all individuals to provide insurance to each other.[23] Of course, such an insurance scheme excludes permanently disabled individuals.

Assuming away exploitation, which I have relegated outside the scope of this paper, the theory presented here nevertheless allows for a simple connection between positive and normative analysis: while a positive analysis reveals norms dictated by the cooperation opportunities currently exploited, a normative analysis highlights new norms, that could be set-up in order to exploit additional cooperation opportunities. The solution to the vulnerability issue lies precisely in the recognition that disability is as much an individual as a societal characteristic. The fact that a person has no way of contributing to the current functioning of society does not mean that there is not a social arrangement in which he could contribute. Defending that a person may never, independently of the social context, contribute to the society's output can not be rationally justified, for the absence of proof does not constitute a proof of the absence. This view, in line with Anderson (1999, p. 321)'s equality in "effective freedom to achieve functioning that are part of citizenship", fits nicely with the lexical ambivalence of the word dignity, that characterises both the rights that a person may be granted in society and his ability to uphold the associated duties. It highlights the danger of an *is-ought* fallacy concerning individual abilities and insists on the importance

[23] In the present framework, R, R_1 and R_2 can be interpreted as certainty equivalents. In the case where agent 1 knows that, with some probability p, agent 2 will produce nothing, then he expects that with this probability p his outside payment R_1 will be equal to R. As long as there is a possibility for agent 2 to contribute in the future, agent 1 may find it worth cooperating. The case where agent 1 has the same probability of being incapacitated as agent 2 is a case of perfect symmetry of the certainty equivalents $R_1 = R_2$.

of disentangling positive and normative statements.

This separation between positive an normative statements is also interesting in itself. Importantly, the equality conditions that allow cooperation and that define the norms of equality entail no inter-personal comparison of utility since they are derived from a comparison of each individual's own gain in different scenarios. It is not the purpose of this paper to discuss the pros and cons of utilitarianism but if we accept, with utilitarians, that distributional choices can be summarised by the optimisation of a weighted sum of utilities, then the weights that can be recovered from the empirical distribution of goods[24] are the direct reflect of the cooperations that actually take place in the economy. Determining which weights should be used, or how weights should be changed, in contrast, amounts to determining which currently in-exploited cooperation opportunities can be set-up to improve the welfare of all.

Imperfect information, status quo biais, time-inconsistent preferences and other market imperfections may prevent individuals from identifying mutually beneficial cooperation opportunities. In such a case, the associated equality constraints fail to be enforced, while it should. The normative content of this affirmation is justified by the Pareto improvement that cooperation allows and the only judgements of value that are made to reach it is that more good is better for both individuals and that rationality should be used. We may discuss the relevance of these two value judgements but, compared with most theories of justice and norms, they are arguably minimal value judgements. The more is better assumption entails little loss of generality, for it is always possible to label the absence of harm, a good.[25] The desirability of rationality can be related to the fact that humans share a common conception of such a means to ends rationality, which makes it an intuitive candidate for backing normative statements. The aim of this paper cannot be to derive a perfect wedge between normative and positive views of norms, but to find normative principles for the organisation of society that derive from the normative principles that individuals apply to themselves, which amounts to saying that we seek to derive moral principles on non-moral grounds.

Of course, identifying Pareto-improving cooperation opportunities is much more complex in the real world than in the two players game described here. The following section nevertheless makes an attempt at generalising the results to an n person society.

[24]The word "good" is used in its broadest sense. It covers anything that is both desirable and scarce. Income, wealth and political power are obvious examples of such goods.

[25]An important situation where the model does not apply that should be acknowledged is when goods are non-rival.

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 generalised mutual cooperation, 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

$$\delta \geq \max_i \frac{g(n) - \phi_i R(n)}{g(n) - R(n)/n + c} \equiv \delta(n, \Phi), \tag{5}$$

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 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

$$\delta \geq \frac{R - R/n}{R - R/n + c} \equiv \delta^*(n). \tag{6}$$

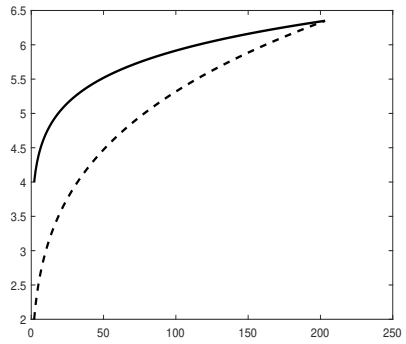
An increase in the number of players raises the relative gain from defection $R - R/n$ (the numerator in Equation 6) by lowering the cooperative share R/n, hence making cooperation harder to achieve, but lowers the payoff in the war of all against all (the denominator), hence favouring cooperation. It is straightforward to see from equation (6) 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.

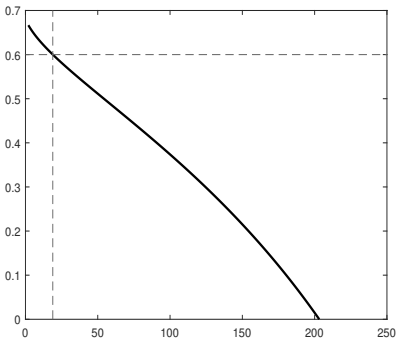
Individuals may therefore cooperate on large scales only if the gain from unilateral deviation $g(n) - R(n)/n$ decreases with n, in which case we have $\delta^*(n)' \leq 0$. Equation (5) shows that, if the deviation payoff $g(n)$ grows at a lower pace than the cooperative payoff $R(n)/n$, then $\delta^*(n)$ may indeed decrease with n. Figure 2 illustrates an example of such a society.[26] 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.[27] 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, 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.

[26] The calibration is reported in Appendix 6.2.

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



(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.[28]

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 in equilibrium. 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

$$\min_{\phi} \delta^*_{GT}(\phi) = \delta^*_{GT}(1/2) \leq \min_{\phi} \delta^*_S(\phi). \tag{7}$$

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 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

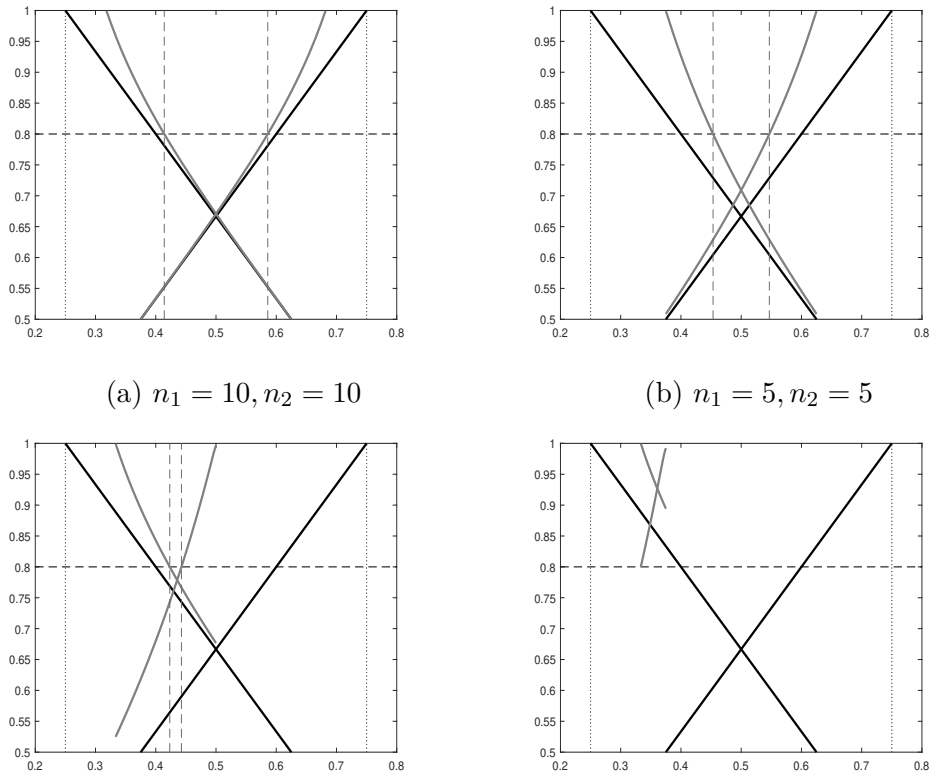
[28]For example, players may take turn playing D while the other plays C. Such cooperating strategies allow to modulate the frequency of each agent playing D and hence receiving the high payoff. Compared to an infinite sequence of (C, C), that gives each players his share ϕ and $1 - \phi$ of the cooperative 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.

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 (7) 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 illustrates 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. A discussion of the asymetric abilities, of the vulnerability issue is then offered. 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 3 summarises the game. The situation (S, S), where

	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 3: 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

$$\frac{e}{R} < \phi < 1 - \frac{e}{R}, \tag{8}$$

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

$$\frac{e}{\delta R} \leq \phi \leq 1 - \frac{e}{\delta R}.$$[29]

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

[29]Since $\delta \leq 1$, cooperation is therefore possible only if condition 8 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.

Anderson, E. S. (1999). What is the point of equality? *Ethics*, 109(2), 287–337.

Aumann, R. & Shapley, L. (1976). Long term competition - a game theoretic analysis. *Mimeo.*

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.

Binmore, K. G. et al. (1994). *Game theory and the social contract: just playing*, volume 2. MIT press.

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.

Dworkin, R. (1981). What is equality? equality of welfare and equality of ressources. *Philosophy & public affairs*, 10, 185–345.

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.

Hampton, J. (1986). *Hobbes and the social contract tradition*. New York: Cambridge University Press.

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. (2009). Why hobbes' state of nature is best modeled by an assurance game. *Utilitas*, 21(3), 297–326.

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

Moehler, M. (2020). *Contractarianism*. Elements in Ethics. Cambridge: Cambridge University Press.

Moreno-Ternero, J. D. & Roemer, J. E. (2008). The veil of ignorance violates priority. *Economics & Philosophy*, 24(2), 233–257.

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

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Roemer, J. E. (1998). Equality of opportunity.

Sen, A. (1980). Equality of what? *The Tanner lecture on human values*, 1, 197–220.

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

Vanderschraaf, P. (2006). War or peace?: A dynamical analysis of anarchy. *Economics and Philosophy*, 22(2), 243.