PARTIAL COMPLIANCE:
SUNDAY SCHOOL MORALITY MEETS GAME THEORY.

Magnus Jiborn

Magnus.jiborn@fil.lu.se

ABSTRACT: There is a striking gap between the moral standards that most of us endorse, and the moral standards that, in practice, we seem able to live up to. This might seem disturbingly hypocritic. However, Wlodek has often suggested in discussions that endorsing high moral standards, a "Sunday school morality" so to speak, can make us behave better than we would otherwise do, even if we do not achieve perfection. I present an argument from evolutionary game theory to support this Sunday school thesis. [1]

[1] This paper began as a coffee room discussion with Wlodek. The paper should be considered as a rough draft for further discussion and, perhaps, a first step to a joint paper.

Printed from: *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz.* Ed. T. Rønnow-Rasmussen, B. Petersson, J. Josefsson & D. Egonsson, 2007. www.fil.lu.se/hommageawlodek

1. Introduction

There is a striking gap between the moral standards that most of us endorse, and the moral standards that, in practice, we seem able to live up to. In words, we subscribe to norms of honesty, cooperation and unselfishness, although we all know very well that these norms will frequently violated and that we will, occasionally, be tempted to violate them ourselves. Honesty might well be a good rule of thumb, but it is also obvious that there are times when we find it in our interest to make exceptions.

If such apparent hypocrisy seems morally unsavoury, there are two alternative ways to deal with it; either we could adjust our moral standards to fit with the observed behaviour, or we could increase efforts to improve compliance until it satisfies our moral ambitions.

But perhaps we should do neither? In discussions over the relation between norms and rationality, Wlodek has often suggested that subscribing to high moral standards, a "Sunday school morality" so to speak, can be a useful way to make us behave better than we would otherwise do, even if we cannot hope to achieve perfection.

In this paper, I shall present an argument from evolutionary game theory to support this Sunday school thesis. I will argue that there are cases where partial compliance is the best we can hope for and where, hence, the best we can do is to accept a certain level of hypocrisy. Morality in these cases might, borrowing a formulation from the political scientist Stephen D. Krasner, be characterized as a system of "organized hypocrisy"[2].

I will consider two different cases. The first one is based on a standard, iterated Prisoner's Dilemma. It is well known that reciprocal cooperation, given certain assumptions, can be an equilibrium in this kind of game, but not a stable one; it is vulnerable to random drift by unconditional cooperation, which might, in turn, be invaded by defection. I show that reciprocal cooperation can be stabilized by the continuous influx of a small amount of mutant defectors.

The second case is based a similar game, but with indirect, community enforcement instead of direct reciprocity between the same two players (Kandori 1992; Nowak and Sigmund 2005). Whereas standard models of community enforcement presuppose that interactions are transparent – i.e. that the behaviour of identifiable players can be monitored by the community – I will consider a case with only partial transparency.

In the standard models, with full transparency, indirect reciprocity is vulnerable to the same weakness as direct reciprocity; when everyone cooperates all the time, there are no evolutionary barriers against undiscriminating cooperation. With only partial transparency, players might sometimes get away with defection. I will show that allowing for a certain level of rational defection in a population provides incentives for vigilance and turns reciprocal cooperation (with some tendency to cheat) into a stable equilibrium.

[2] Krasner uses the term "organized hypocrisy" to characterize international norms of state sovereignty (Krasner 1999).

A certain level of cheating within a community serves to drive out the unduly meek and keep reciprocal cooperators on the edge.

Finally, I discuss the implications of these results for the relation between moral norms and individual rationality.

2. Background and method

There is an influential tradition, inspired by David Hume, that seeks to account for moral norms in terms of conventions. Norms, on this account, are rules of behaviour that people comply with on the expectation that others will comply as well. In a well known passage, Hume explains property rules in this way:

> I observe that it will be in my interest to leave another in the possession of his goods, *provided* he will act in the same manner with regard to me. He is sensible of a like interest in the regulation of his conduct. When this common sense of interest is mutually express'd, and is known to both, it produces a suitable resolution and behaviour. (Hume 1978: 490)

Modern writers in the Humean tradition have often used evolutionary game theory to account for the process by which moral norms may emerge and become established in a society(Sugden 1986; Bicchieri 1990; 1993; Binmore 1994; Skyrms 1996; Binmore 1998; Skyrms 2004). The rationale behind this approach is this:

Norms are thought of as behavioural regularities that may emerge among a group of people in response to some recurrent problems of social interaction. Game theory offers a convenient way to represent the structure of such interaction problems. Classical game theory, however, operates with the assumption that players are ideally rational and have access to perfect information about the game they are playing. This is hardly a realistic assumption.

More realistically, we may assume that players who are repeatedly exposed to the same type of problem can learn from experience. The precise form of the learning mechanism is not very important; it could be a simple process of trial and error or a process of imitation. Given that players are more likely to abandon strategies that have proved less successful, and/or more likely to adopt strategies that have proved more successful, learning will result in an adaptive process that is in many respects similar to the process of natural selection in biological evolution. Evolutionary game theory supplies analytical tools to model the outcome of such a process of adaptation.

The central solution concept of classical game theory is the *Nash equilibrium*. A Nash equilibrium obtains when the strategy chosen by each player is a best response to the strategies chosen by all others. A *strict* Nash equilibrium obtains when the strategy of each is the unique best response to the strategy profile of the other players. That is, given what everyone else does, each player would be strictly worse off by unilaterally choosing differently.

In evolutionary game theory, an equilibrium is a population state rather than the outcome of a single game. Each player is thought tho be pre-programmed for some strategy, and the strategy profile of the population is updated in accordance with average payoffs to different strategies. In equilibrium, the strategy profile of the population is at rest, since all strategies that are present in the population earn the same payoff. An equilibrium is stable if the population, after any small disturbance, returns to the equilibrium state.

On the evolutionary account, a convention is a stable equilibrium in a game with more than one equilibrium (Sugden 1986). To account for property rules or norms of cooperation, according to this tradition, we must show that the behaviour prescribed by these norms constitute stable equilibria in the relevant kind of game, given a plausible adaptive dynamics.

A stable equilibrium is often identified with an *Evolutionary stable strategy*, or ESS (Maynard Smith 1997). An ESS is a strategy such that, once nearly everyone in a population follows it, it cannot be invaded by any "mutant" strategy[3]. Formally, a strategy is an ESS if it satisfies the following conditions. Let $u(s, s')$ be the average payoff to strategy s when matched against strategy s'. Then s^* is an ESS iff, for any possible mutant strategy s, either :

(I) $u(s^*, s^*) > u(s, s^*)$, or

(II) $u(s^*, s^*) = u(s, s^*)$ and $u(s^*, s) > u(s, s)$

However, in section 3, I will suggest that there can be stable equilibria that are not ESS. In particular, I will argue that allowing for continuous "experimentation" with non-equilibrium strategies – that is, assuming a constant small influx of mutants to the population – can stabilize a cooperative equilibrium that is not an ESS.

[3] The term "mutant" is, of course, loaded with biological connotations. In a learning dynamics, we may think of mutation as the result of experimentation; occasionally some player will adopt a new strategy at random; if it does well it will gain new followers in subsequent rounds, otherwise it will soon disappear.

3. Prisoner’s dilemma

We begin by considering a standard Prisoner’s Dilemma game with the following payoff matrix:

	C	D
C	3, 3	0, 4
D	4, 0	1, 1

Figure 1. Prisoner’s Dilemma

Each player chooses independently whether to cooperate (C) or defect (D). Each player cares only about maximizing her own payoff. For each player, defection dominates cooperation, i.e. regardless of what the other player does, each gains by choosing D rather than C. On the other hand, if both defect, they will both be worse off than if they both cooperate.

Hence the dilemma. Individually rational behaviour may well lead to collective disaster.

But suppose the game is played repeatedly an indefinite number of times between the same two players. Players might then condition their choice of action on the history of the game. For example, after the first round, each may choose to cooperate if and only if the other cooperated in the previous round. Or to cooperate as long as the other cooperates, but to defect forever if the other player defects even once. But they might, of course, also choose to always defect or to always cooperate, regardless of what the other player did. There are innumerable possible strategies in this kind of iterated game.

According to the so called Folk Theorem, cooperation based on reciprocity can be viable in such an indefinitely repeated PD, given that the prospect of future games is sufficiently important. Each player must then take the possible effects of her current behaviour on the future behaviour of her opponent into account. The "shadow of the future" will affect current incentives.

The most simple and well known reciprocal strategy in the iterated PD is the so called *Tit for tat*, *TFT*, that was made famous through Robert Axelrod's classical book *The Evolution of Cooperation.* Axelrod showed that, even if *TFT* is not the best response to every opponent strategy or every environment, it is robust in the sense that it does well in many social environments, and that it should have a good chance in the evolutionary competition. If a player plays *TFT*, there is no use for others to try to exploit her; defecting against *TFT* is immediately punished in the following round. The best response when playing against *TFT* is to cooperate, and the best strategy to use in a population that is committed to *TFT* is also to cooperate. A population of *TFT* players cannot be invaded by defectors, since defectors will do strictly worse in that environment than the *TFT* players themselves.

Axelrod claimed that *TFT* is a stable equilibrium, an ESS, when played under an evolutionary dynamics(Axelrod 1990)[4]. Unfortunately, however, that claim has been shown not to hold for closer scrutiny (Binmore 1994). It is true

[4] Axelrod uses the term ”collectively stable”, but the definition of collective stability is very similar to that of evolutionary stability.

that *TFT*, once established in a population, can withstand invasion by non-cooperating strategies, such as *Always defect, AD*. But there will be no evolutionary pressure against less discriminating strategies, such as *Always cooperate*, *AC*. As long as everyone cooperates all the time, there is no difference in behaviour or payoffs between conditional and unconditional cooperation. Hence, *AC* might enter the population by random drift, and once it has become sufficiently frequent, the population is an easy prey for *AD* to invade.

To put it more formally, $u(TFT, TFT) = u(AC, TFT)$ and $u(TFT, AC) = u(AC, AC)$. Evolutionary stability would require either $u(TFT, TFT) > u(AC, TFT)$ or $u(TFT, TFT) = u(AC, TFT)$ and $u(TFT, AC) > u(AC, AC)$. Hence, *TFT* does not satisfy any of the two criteria for an ESS. *TFT* is an equilibrium but not a stable one.

Neither is *AD*. Once established, *AD* does strictly better than any mutant strategy that begins by cooperating. Hence, it can withstand invasion from cooperative strategies such as *TFT*. But it is not protected against a more cautious version of *TFT*, for example one that starts by defecting in the first round and then only cooperates if the other cooperates first. Let us call this strategy *CTFT*, *Cautious tit for tat*. As long as everyone defects all the time, there is no difference with regard to either behaviour or payoffs between *CTFT* and *AD*. Hence there is no evolutionary pressure against *CTFT* which can enter by random drift.

Once *CTFT* has become sufficiently frequent, the population becomes vulnerable to invasion by either *TFT* or some modified version of *TFT* that can push the population towards universal cooperation. When *CTFT* meets *TFT*,

TFT begins by cooperating, whereas *CTFT* does not. Hence in the second round, *TFT* punishes its opponent by defecting, whereas *CTFT* cooperates. In the third round, then, *TFT* returns to cooperation, whereas *CTFT* defects. Thus, *u*(*TFT*, *CTFT*) =, which is larger than *u*(*CTFT*, *CTFT*) = 1. Hence, *TFT* can invade *CTFT*. In fact, since *TFT* does only slightly worse than *AD* against *AD*, a rather small proportion of *CTFT* is enough to make invasion by *TFT* possible.

There is a modified, somewhat more forgiving, version of *TFT*, that begins by cooperating twice, and only after that does exactly what the opponent did in the previous round. Let us call this strategy. *TFT** would do even better than *TFT* against *CTFT*. On the other hand, it would do worse against *AD*. Which of these strategies will gain the upper hand probably depends on the relative proportions of *AD* and *CTFT* when the invasion begins. The question would have to be tested in computer simulations. Any case, it seems clear that *CTFT* can enter *AD* by random drift, and that, once sufficiently frequent, *CTFT* can be invaded by either *TFT* or some version of *TFT* that can establish cooperation. Since cooperation is not stable either, the population is likely to oscillate between cooperative and non-cooperative states.

The conclusion, so far, is that neither universal cooperation nor universal defection are stable equilibria. Nor is there a stable mix between cooperative, such as *TFT* or *AC* and non-cooperative strategies such as *AD*. This is also easily demonstrated.

Suppose, as a *reductio*, that there is some probability mixture s^* =

(*pAD*, *qAC*, (1-p-q)*TFT*) – where p and q are proportions ($0<p<1$, $0<q<1$ and $0<p+q<1$) – between *AC*, *AD* and *TFT* that amounts to a mixed strategy ESS. In a mixed strategy ESS, every component pure strategy, and hence every mixed strategy that is a probability distribution over the component pure strategies, must earn the same average payoff. It must be the case, then, that $u(AD, s^*) = u(s^*, s^*)$. Thus, if s^* is an ESS, it must be the case that $u(s^*, AD) > u(AD, AD)$. But obviously, $u(AD, AD) > u(s^*, AD)$. Hence, s^* cannot be an ESS.

But there is another possibility.

Let us consider the concept of mutation that plays an essential role in the evolutionary analysis. An ESS is defined as a strategy that can withstand invasion by any mutant strategy. Hence, to motivate the ESS concept, we must assume that there is some mutation mechanism. In biological evolution mutations are thought to occur due to recombination or errors in copying DNA material. The idea in cultural evolution is that players will sometimes "experiment" by trying a new strategy. Such experiments could be conscious, or they could be due to mistakes.

In the standard analysis, the mutation frequency is thought to be very low; it is not supposed to affect the population state and hence not the average payoffs to different strategies that occur in the differences and equations that define the ESS concept. But suppose that mutations are more frequent. In a large population, we may assume, there will always be some players who experiment. Hence, we may assume there will always be a small proportion of players currently playing, for example, *AD*. If the population is in the *TFT* equilibrium, these experimenters will be punished and quickly return to the equilibrium orthodoxy. But they will be replaced by other experimenters. The effect is that, although *AD* does strictly worse than *TFT*, it never disappears completely from the population; there is always a positive probability of being paired with an *AD* player.

The effect of this assumption is that *AC* will do strictly worse than *TFT*. Hence, *AC* will no longer be able to take over by random drift, and the *TFT* equilibrium will be stabilized. Of course, players might experiment with *AC* as well. Hence, this strategy will never completely disappear either. But, since *AC* players do strictly worse than the population average, they will not increase.

Or, to be more precise, whether or not the proportion of *AC* will grow, depends on the relation between the selection pressure provided by the adaptive dynamics and the rate of mutation. Here, I have not specified the dynamics, that is, the relation between growth rates and payoffs. I have only assumed, rather vaguely, that "strategies that earn more than average become more frequent, whereas strategies that earn less become less frequent". That is, I have assumed a payoff positive dynamics, which all we need to define the concept of ESS.

It is quite possible, or even very probable, that there exist some combinations of selection dynamics and mutation rates that would have the population stabilize on a certain mixture between reciprocal and unconditional cooperation and unconditional defection. This issue should be further analyzed by computer simulation.

The main conclusion however, is that the continuous presence of a small proportion of non-cooperation might serve to stabilize an otherwise unstable cooperative equilibrium. Hence, to sustain cooperation in a case like this, a community should punish defectors, but avoid pushing down the rate of experimentation with non-cooperative strategies too far. Occasional defections help to keep reciprocal cooperation on the edge.

4. Community enforcement with partial transparency.

It has been shown by a number of writers that the Folk Theorem for iterated games can be extended to situations where players interact repeatedly, not with the same opponent, but with players from the same community. Even if each pair meets only once, cooperation can be sustained if information about previous behaviour can be transmitted within the community. The idea is that each player builds up a reputation, based on her behaviour in previous games, and that a player's choice of action in a certain game might be conditioned on the reputation of her opponent(Kandori 1992; Nowak and Sigmund 2005).

I will consider a game that is similar in structure to the standard PD, but where players alternate between two roles – whenever two players meet, one of them has the opportunity to help the other by conferring some benefit b to the other, at some cost c to herself.[5] We assume that $c<b$; thus giving and receiving is better for each than not giving and not receiving.

Let us suppose that pairs of players are drawn at random from a population, one plays the role of potential donor, the other the role of potential receiver. If the potential donor donates, the receiver gets b and the donor $–c$. If the potential donor refuses, they both get 0. After such a meeting, players return

[5] I here follow the model suggested by (Nowak and Sigmund 2005)

with their payoffs to the population and never meet again. The receiver, thus, will never have the opportunity to pay back or retaliate.

If interactions are anonymous, it is obvious that refusing to donate is the only possible equilibrium; regardless of what others in the population do, a player always benefits by not donating. However, this situation changes if interactions are not anonymous and if information about previous behaviour can be transmitted within the community.

Suppose, for simplicity, that each player carries one of two alternative labels, either *Innocent* or *Guilty* and that these labels are updated after every round according to how the player acted in that round. Before deciding whether to donate or not, a player in the role of potential donator may check the reputation label of her opponent, and to donate if and only if the other is innocent (or guilty – but I will disregard this alternative here). Let us call this strategy *Conditional cooperation, CC.* Cooperation here means to donate if in the position of potential donator..

Each player may of course also choose to disregard labels and *Cooperate*, *C*, without conditions. Or to *Defect*, *D,* regardless of the other player's reputation, where defection means to refuse to donate when in the position of potential donator. For simplicity, I will only consider these three alternatives at the moment.

There are some different possible rules for assigning labels. For example, a player could be labelled *Innocent* if and only if she gave in her previous . Or, alternatively, she could be labelled *Guilty* only if she failed to cooperate with a player labelled *Innocent*. Or, a third possibility, *Guilty* only if she either failed to cooperate with an innocent player, or failed to defect against a guilty player.

In a detailed model, the choice of labelling rule should probably be treated as the result of a process of co-evolution of strategy and labelling. I will here simply assume that the second rule is in use. The first rule might seem appealing because of its simplicity, but on the other hand it seems strange that players who punish guilty players would thereby become guilty – and perhaps punished – themselves (see (Nowak and Sigmund 2005) for a discussion of this issue). The third rule is harsh on the meek and thereby favours conditional over unconditional cooperation. I have chosen the second rule as a middle way. I believe, however, that the main argument could be stated equally well with the third rule. The first rule, on the other hand, has somewhat different implications.

How information is transmitted and whether it is reliable is obviously a key factor. In Kandori's model the information system is treated as exogenous and reliable. I will follow him in this, although I believe that in a more detailed and realistic model, we should have to consider both the problems about reliability and the cost of information processing. This is a subject for further analysis, but one I will not carry out here.

Given a payoff positive selection dynamics, it is easily seen that *CC* is an equilibrium under these assumptions. When everyone plays *CC*, no one can gain by unilaterally choosing a different strategy. *AC* yields the same payoff as *CC* whereas playing *AD* results in being labelled *Guilty* and hence in being punished by not receiving any future benefits.

To illustrate this, let us consider two consecutive rounds of the game. For simplicity, we assume that if a player is given the role of potential donor in one round, she will be potential receiver in the next.[6] Suppose that everyone else plays CC. The expected payoff to a player currently in the role of potential donor is:

	Round 1	Round 2	Total
CC	*-c*	*b*	*b-c*
AC	*-c*	*b*	*b-c*
AD	0	0	0

Since $c<b$ it follows that *AD* earns strictly less than both *CC* and *AC* in an environment dominated by conditional cooperation.

CC is not an *ESS*, however, for the same reason that *TFT* is not an *ESS* in the iterated game discussed in the previous section; since *AC* earns the same as *CC*, it can enter by random drift, and eventually make the population vulnerable to invasion by *AD*.

However, *AD* is not an *ESS* either; in a population where everyone plays *AD*, everyone carries the label *Guilty*, and hence *CC* always refuses to cooperate. *CC*, thus earns the same payoff as *AD* and can enter by random drift. Of course, there will be no cooperation in such a state, but if conditional cooperators, by random drift, become numerous enough, it is possible that a mutant unconditional cooperator might eventually start a chain reaction that tips the population state back to conditional cooperation.

[6] This assumption does not affect the results of the analysis in any substantial way.

The main message here, however, is that while cooperation based on community enforcement, or indirect reciprocity, can be an equilibrium it is not stable. Continued cooperation requires the population to be vigilant and discriminating, but when everyone cooperates, there are no evolutionary barriers against meekness. And widespread meekness will eventually destroy the conditions for the cooperative equilibrium.

It is possible that an argument similar to the one presented in the previous section, based on mutation rates, could be developed for community enforcement as well. However, I will here take a slightly different approach.

Indirect reciprocity presupposes transparency. Defection can be punished only if it is known who the defector is.[7] Community enforcement rest on the idea that behaviour affects reputation. However, if players interact with many others within a large community, it is likely that some of those interactions will be anonymous. If games are played anonymously, reputation looses its bite. With complete anonymity, unconditional defection is the dominant strategy and the only equilibrium.

Games need not be either completely transparent or completely anonymous, however. If a player defects against an innocent opponent, there could be a certain probability, rather than complete certainty, that this will result in that player having a bad reputation in the next round. Whether or not conditional cooperation is an (unstable) equilibrium depends on the level of transparency.

[7] This is a truth with modification. Kandori has shown that community enforcement based on collective punishment is a Nash-equilibrium (Kandori 1992). It is a rather fragile equilibrium however; if players can make errors it is unlikely that there will be much cooperation going on in the long run.

If the level of transparency is below a certain threshold, universal defection is the only equilibrium.

In a realistic model, however, the level of transparency can be supposed to vary, so that some rounds are played with a high level of transparency, whereas others are played with almost complete anonymity. With this assumption, a whole range of new possible strategies emerge. Players may now choose to cooperate on condition that their opponent is innocent *and* on condition that the level of transparency is above a certain level. Let CC_x be strategy of cooperating if and only if the opponent is innocent *and* the level of transparency is at least x. Likewise, let C_x be the strategy of cooperating regardless of the other players label, but only if the level of transparency is above x.

Of course players may also choose to disregard the level of transparency and play a straightforward strategy C, D or CC.

Let l_i be the transparency level of a round i and let p_x be the probability that the transparency level is above x, where $0 \leq x \leq 1$. Now, the prospects for conditional cooperation in this setting will obviously depend on a number of parameters: the relation between c and b and the distribution of different transparency level. Is near anonymity common or uncommon?

What I intend to show, however, is that if there exists some number m, $0 < m < 1$, such that $mp_m = c/b$, then CC_m is a stable equilibrium.

Let us consider the expected payoffs to different possible strategies in an environment where everyone plays CC_m. We check the payoffs in two consecutive rounds for a player who begins in the position of donator, assuming for simplicity that a player who is donator in this round will be receiver in the next. In the first round, the potential receiver will be either guilty or innocent. If the potential receiver is guilty, a player using CC_m will refuse to give in that round, hence she will have payoff 0 . In the second round, she will still be innocent. Whether or not she will receive a benefit depends on whether the transparency level, l_2, in the second round is at least m. Her payoff in the second round, then, if she was matched against a guilty player in the first round, will be $p_m b$.

Suppose the CC_m player is matched against an innocent player in the first round. If $l_1 \geq m$ she will donate, and carry a cost of $-c$ in the first round. In the second round she will be innocent and receive $p_m b$. If, on the other hand, $l_1 < m$, she will refuse to donate and have 0 in that round. In the second round, there will be a certain probability $\pi < m$, that she will be labelled guilty and have 0, and a probability $(1-\pi)$ that she will still be labelled innocent and have $p_m b$.

We may summarize the expected payoff, $u(CC_m, CC_m)$ in the following way:

CC_m

		Receiver in first round is	
		Guilty	Innocent
$l_1 \geq m$	$p_m b$	$p_m b$ -c	
	$l_1 < m$	$p_m b$	$(1-\pi)\ p_m b = p_m b - \pi\, p_m b$

Compare this to the strategy *C*, which cooperates on every occasion when in the role of potential donor. Hence it gets $p_m b - c$ all the time. But $p_m b - c$ is less than $p_m b$ and, since $\pi < m$, it is also less than $p_m b - \pi p_m bk$. The strategy *C*, thus, never gets more than CC_m and often less. Hence it does strictly worse against CC_m than CC_m does itself.

Compare also with *D*, which defects regardless of the reputation of the other. It earns the same payoff as CC_m when matched against a guilty player in the first round. It also earns the same as CC_m when $l_1 < m$. However, when $l_1 > m$, CC_m cooperates and gets $p_m b - c$, whereas *D* defects and gets $p_m b - \pi^* p_m b$ where $\pi^* > m$. Since $\pi^* > m$ it follows that $\pi^* p_m b > c$. Hence, *D* never earns more than CC_m and sometimes less. It too, then, does strictly worse against CC_m than CC_m does itself.

Now, is it possible that there could there be a strategy that does better against CC_m than CC_m does itself? Let us try to answer this question by considering how a strategy could *differ* from CC_m.

First, it could behave differently by sometimes cooperating with a guilty player. That is hardly a recipe for success, however. Cooperating with a guilty player will yield $p_m b - c$ instead of $p_m b$ so it is a sure loss.

Second, it could differ by sometimes defecting against an innocent player even if $l_1 > m$. However, as we have seen, that yields $p_m b - \pi^* p_m b$ when CC_m earns $p_m b - c$, and since $\pi^* > m$ it follows that $p_m b - \pi^* p_m b < p_m b - c$.

Third, it could differ by sometimes cooperating even if $l_1 < m$. It will then have $p_m b - c$, whereas CC_m gets $p_m b - \pi p_m b$. But, as we have seen, since $\pi < m$, $p_m b - c < p_m b - \pi p_m b$.

The only way that a strategy could differ from CC_m in an environment where everyone plays CC_m and still not do strictly worse than CC_m itself, is by not cooperating when $l_1=m$. It then earns precisely the same as CC_m. Hence, there is a slight modification of CC_m that acts exactly as CC_m but only cooperates if the level of transparency strictly exceeds m, instead of when it is at least m. These two, nearly identical, strategies earn the same. Any other strategy will earn strictly less. The equilibrium, thus, could consist in a random mix of CC_m and its close relative, but this equilibrium cannot be invaded by any other strategy. It is stable.

This is a rather striking result. Introducing an opportunity to sometimes cheat and have good chance of getting away with it, apparently serves to stabilize an otherwise unstable equilibrium of conditional cooperation based on indirect reciprocity. Again, it seems that the presence of a certain, hopefully small, fraction of defection is necessary to keep a conditionally cooperative population on the edge and protect it against the dangers of meekness.

5. Discussion

David Hume is among the moral philosophers who have claimed that morality in some sense must rest on rationality:

> What theory of morals can ever serve any useful purpose, unless it can be shown that all the duties it recommends are also the true interest of each individual?(Hume 1975: 280)

David Gauthier, on the other hand, rejects this view of the relation between morality and rationality:

> David Hume, who asked this question, seems mistaken; such a theory would be too useful. Were duty no more than interest, morals would be superfluous. (Gauthier 1986:1)

Gauthier claims that a norm, in order to be a moral norm, must sometimes require us to act in ways that are contrary to our direct self interest. Gauthier therefore rejects norms that are based on direct or indirect reciprocity as truly moral norms. Such norms appeal to nothing else than direct self interest and, thus, makes morality superfluous. Gauthier's well known solution to this seeming dilemma for a theory that attempts to establish morality on rational agreement is the idea of rational commitment.

However, given the result presented above, the relation between moral norms and rationality seems to be more complex than Gauthier thinks. The community enforcement norm in the case discussed above must require that players *always* cooperate with those who are innocent. It is collectively rational for a group to subscribe to such a norm, since it allows them to cooperate for mutual benefit. And it is also *generally* in the interest of each individual to act on that norm. But not always. Not if the risk of being punished is sufficiently small. The norm, thus, requires more than is supported by direct individual rationality. The system will be one of organized hypocrisy in this sense. Everyone subscribes to a norm that requires that they always cooperate with other cooperators, but most members will cheat whenever they think they can get away with it.

And this is good! It is because they cheat that the level of cooperation can be sustained. Perhaps, one might think, the norm should be adjusted so that people are not punished in cases where everyone cheats. After all, cheating is, in a sense, socially useful. So why punish it?

However, it is easy to show that lowering the norm so that it only prohibits cheating when the risk of being caught is above *m* destabilizes the cooperative equilibrium. There will then no longer be any guilty individuals in the population, which means the difference between unconditional and conditional cooperation disappears. A strategy such as C_m, that cooperates regardless of the other player's reputation, given that the transparency level is at least *m*, will do as well as CC_m. It can therefore enter by random drift and, when sufficiently frequent, enable defection to invade and take over.

But likewise, trying to improve behaviour by increasing transparency may also jeopardize cooperation. Perfect transparency and full compliance is not desirable. If it was achieved, it would destabilize cooperation by reducing the incentive to be vigilant, and eventually open up for invasion by defection.

The conclusion is that organized moral hypocrisy, with stringent norms but only partial compliance, might be the best we can hope for when it comes to reciprocal cooperation. That means we should perhaps continue to pretend to be better than we are, and be morally upset when people are caught cheating in ways that we gladly do ourselves whenever we think we can get away with it. Perfect compliance can perhaps never be achieved. Partial compliance is possible, but only if those who are not without sin are prepared to throw the first stone.

References

Axelrod, R. (1990), *The evolution of cooperation*. London: Penguin.

Bicchieri, C. (1990). "Norms of Cooperation." *Ethics* **100**(4): 838-861.

Bicchieri, C. (1993), *Rationality and Coordination*. New York: Cambridge University Press.

Binmore, K. (1994), *Playing Fair: Game theory and the social contract, vol 1.*. Cambridge, Mass.: The MIT Press.

Binmore, K. (1998), *Just Playing: Game theory and the social contract, vol 2.*. Cambridge, Mass.: The MIT Press.

Gauthier, D. (1986), *Morals by Agreement*. Paperback edition, 1988 ed. Oxford: Oxford University Press.

Hume, D. (1975), *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Edited by L.A. Selby Bigge. 3rd ed. ed. Oxford: Clarendon Press, (originally published 1777).

Hume, D. (1978), *A Treatise of Human Nature*. Edited by L.A. Selby Bigge. 2nd ed. ed. Oxford: Oxford University Press, (originally published 1739).

Kandori, M. (1992). "Social Norms and Community Enforcement." *The Review of Economic Studies* **59**(1): 63-80.

Krasner, S. D. (1999), *Sovereignty: Organized Hypocrisy*. Princeton: Princeton University Press.

Maynard Smith, J. (1997), *Evolution and the Theory of Games*. Paperback ed. Cambridge: Cambridge University Press, (originally published 1982).

Nowak, M. A. and K. Sigmund (2005). "Evolution of indirect reciprocity." *Nature* **437**(7063): 1291-1299.

Skyrms, B. (1996), *Evolution of the Social Contract*. New York: Cambridge University Press.

Skyrms, B. (2004), *The Stag Hunt and the Evolution of Social Structure*. New York: Cambridge University Press.

Sugden, R. (1986), *The Economics of Rights, Cooperation and Welfare*. 2nd edition, 2004 ed: Palgrave Macmillan, (originally published 1986).